Dijlah Journal of Engineering Science (DJES) Vol. 3, No. 2, June, 2026, pp.39-48
ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 146

Bias and Fairness in Ensemble Learning: A Review

Maan Y Anad Alsaleem, Omar Shakir Hasan [2], Yahya Albugg [3]
[1,2] Directorate of Education in Nineveh
[3] Northern Technical University(NTU)

Article Info

Article history:

Received Dec., 08, 2025
Revised Feb.,7, 2026
Accepted Mar.,5, 2026

Keywords:

Machine learning
Ensemble learning
Bias
Fairness
Computational justice

ABSTRACT

Fairness and Bias have become important topics related to Machine Learning (ML) as well as the impact they will have on the Performance of an AI System in the real world. This article reviews the available Ensemble Learning Techniques and discusses Fairness-Aware Ensemble Learning by reviewing three general Interventions (Pre-Processing, In-Processing, Post-Processing). Furthermore, we will look at the most popular Fairness Metrics, such as Demographic Parity, Equal Opportunity, and Equal Probability, and the role of sensitive features when evaluating potential Algorithmic Discrimination and illustrate that Fairness-Aware Ensemble Learning provides managers with significant opportunities to address Bias; however, there are still several challenges faced in achieving a balance between Prediction Accuracy and Fairness, as well as a Lack of Interpretability and Multiple Evaluation Metrics available. Our future directions will examine Multi-Objective Frameworks that incorporate Fairness, Accuracy, and Transparency into the design of Reliable/Fair Ensemble Learning Systems.

Corresponding Author:

Omar Shakir Hasan
Directorate of Education in Nineveh
Al-Hadbaa District, Mosul, Iraq
Email: omarshakir06@gmail.com

1. INTRODUCTION

With AI, humans' cognitive abilities may be replicated through the use of technology in ways similar to how we do so traditionally through machines. Machine learning (ML) is one component of AI that concentrates solely on ways for computers to learn, adapt, grow, and enhance themselves through the information obtained by human experience or through collecting massive amounts of data where computer systems can analyze and identify patterns on their own. ML employs both statistical methodology and algorithms to perform these functions by analyzing historical information for its application (Bennett & Babb, 2024; Caiazza & Mazzola, 2024; Huang et al., 2025). With the abundance of data and improved computing power, machine learning techniques have made significant progress and are widely used in many aspects closely related to people's lives (Huang, Lyu, & Wu, 2024a). Recently, a type of machine learning called ensemble learning has emerged (Raftopoulos, Ntoutsi, & Iosifidis, 2025). These are meta-learning algorithms that combine multiple machine learning techniques to build a single predictive model (Gohar, Biswas, & Rajan, 2022). The basic principle of ensemble modeling is to combine several weaker models to form a single, more accurate model (Feffer & Hirzel, 2022). Most machine learning errors arise from noise, bias, and variance. Therefore, ensemble learning methods aim to reduce these factors (Zhang et al., 2022a). Ensemble learning methods can be divided into two types: bagging and boosting (Iosifidis, Roy, & Ntoutsi, 2022). Parallel ensemble learning (bagging) methods combine the results of multiple machine learning algorithms in parallel. By exploiting the independence of algorithms, they can reduce prediction errors through averaging, voting, or the use of a super-learner (Raftopoulos, Papamitsiou, & Giannakos, 2025). In practice, random forest is a commonly used algorithm for parallel ensemble learning and is often used for feature selection (Arévalo-Cordovilla et al., 2025). Sequential ensemble learning (boosting) methods concentrate multiple machine learning algorithms in succession, with the latter relying on the results of previous algorithms (Iosifidis, Roy, & Ntoutsi, 2022). For example,

misclassified examples are given higher weight, allowing subsequent algorithms to focus on these mispredictions and improve overall performance (Ni, Han, Chen, & Zhang, 2024). Ensemble learning techniques have made significant progress and have been widely used in many aspects closely related to people's lives (Zhang, Wen, & Yao, 2024). The nature and technical characteristics of machine learning are influenced by the nature of the data, and its predictions and decisions will produce a certain degree of bias or unfairness (Hu, Wang, Zhang, & Sun, 2024). This issue has gradually attracted the attention of scientific research, industry practitioners, and the public (Yang et al., 2023). In the prediction and decision-making process, fairness means the absence of bias, favoritism, discrimination, or unfairness based on the inherent or acquired characteristics of individuals or groups (Ford, 2025). Therefore, an unfair algorithm means that its decision is biased against a specific individual or group, resulting in unfair treatment of the individual or group and harming their interests (Singh & Ghosh, 2025). Almost all classification algorithms adopt a "myopic" utility optimization strategy, with machine learning algorithms predicting based on data (Huang, Lyu, & Wu, 2024b). Using data to create an algorithm that predicts whether or not someone has heart disease can be problematic because algorithms that were trained using this form of data may overemphasize the influence of age and therefore produce an unfair advantage for all patients over 65 (Pham et. al., 2025). In addition to the age influence, there are also many other ways to introduce bias into machine learning diagnostic systems, such as socioeconomic status (Huang et al., 2024a). As a point of reference, consider the heart disease database created by the University of California-Irvine machine learning repository, which contains records of many patients in varying age groups. The purpose of this database was to assist researchers in developing machine learning algorithms to intelligently diagnose heart disease. However, research has shown that these algorithms created using this database produce bias due to their interpretation of the records and result in giving all patients aged 65 years and older an unfair advantage (Hasanzadeh et al., 2025). In some resume screening tools, discriminatory behavior toward the sensitive attribute of gender occurs, leading men to gain an advantage over women in the application process (Gohar et al., 2022). Machine learning has been applied in many fields and has had a significant impact on people's work and lives (Huang, Lyu, & Wu, 2024b). However, issues of fairness and bias directly impact the level of community and public trust in AI systems, as well as the implementation and deployment of AI systems (Ford, 2025). This presents a new challenge for research and development of machine learning and ensemble learning applications (Feffer & Hirzel, 2022). Aggregated models may inherit the same biases present in the training set, making the bias more persistent (Zhang et al., 2022b). Some ensemble techniques (such as boosting) give greater weight to difficult examples, which may be associated with a particular population, thus reproducing bias against them (Ni et al., 2024). Furthermore, evaluation metrics such as accuracy highlight the superiority of a model without considering fairness (Yang et al., 2023). Ensemble learning also reduces interpretability, which in turn makes bias detection difficult (Raftopoulos, Ntoutsi, & Iosifidis, 2025). Thus, the aim of this paper is to advance the existing knowledge of bias and fairness within ensemble learning models. The outline of the research is as follows: Section 2 provides an explanation of our review method along with an explanation of why we chose that method; Section 3 covers defining bias in machine learning and proposed metrics for measuring fairness; Section 4 addresses ensemble learning bias and fairness, with a focus on pre-, in-, and post-processing; Section 5 compares the various studies that were reviewed; Section 6 summarizes the most important challenges and research directions; and Section 7 concludes the research with a conclusion.

2 .RESEARCH METHODOLOGY

In order to conduct this review, a methodology was adopted to ensure coverage of the literature on bias and fairness in ensemble learning. Relevant studies were collected from major databases such as IEEE Xplore, SpringerLink, ScienceDirect, ACM Digital Library, Scopus, and arXiv, using combinations of keywords including “ensemble learning and fairness,” “bias mitigation in ensembles,” and “fairness-aware bagging, boosting, or stacking.” The initial search yielded more than 150 studies published between 2020 and 2025, which were then filtered based on inclusion criteria that required explicit discussion of fairness, bias mitigation, or equity in ensemble learning models. Papers that focused on accuracy and did not observe fairness were excluded. After removing duplicate data and examining abstracts, approximately 70 papers were reviewed, and 22 were selected for analysis. These papers were classified into three main groups based on the stage of fairness implementation: pre-treatment approaches, training approaches, and post-treatment approaches. Figure 1 illustrates the classification of research papers.

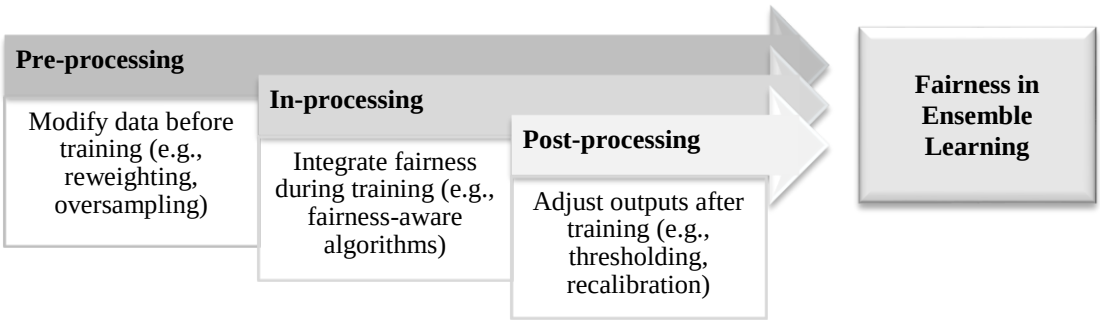

Figure 1. Taxonomy of Fairness in Ensemble Learning

3 .BIAS IN MACHINE LEARNING

Bias is defined as the penchant for unfairness associated with favoring an individual or group (Ferrara, 2023). In computer science, bias can be described as the likelihood that a computer system repeatedly makes the same mistakes and yields 'unfair' results (Yang, Wang, and Li, 2024) (Hanna et al., 2025). This type of bias, which often contains prior information about the past that is a precondition of intelligent behavior, becomes problematic when it arises from aspects of human culture, which are recognized as a cause of harmful behavior (Ferrara, 2023). Bias, on the other hand, is a systematic error that modifies a person's behavior or judgment about another person because of their membership in a group that is generally distinguished by characteristics, such as gender or age (Lindlof & Siegert, 2025). In the realm of artificial intelligence, bias has become defined as inaccurate responses or error-prone decisions because an unintended judgment or bias was introduced into machine learning models (Yang, Wang and Li, 2024). Bias represents a conscious departure from objective reality (Hanna et al., 2025). A model exhibits bias when it has a systematic error that keeps it from estimating the true value of its output (Lindlöf and Siegert, 2025). Figure 2 provides an overview of the types of bias in machine learning (Ferrara, 2023).

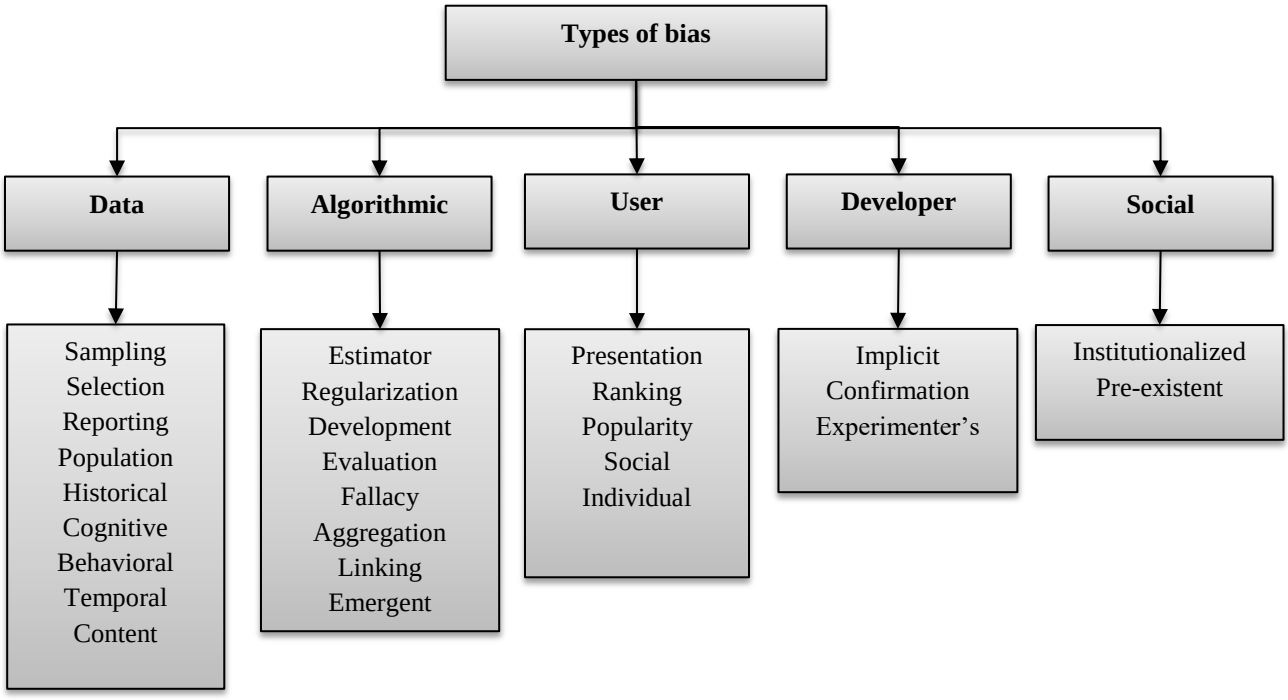

Figure 2. Types of bias in machine learning**.**

3.1 Data Bias

The characteristics of the data used to train models give rise to data bias (Mehrabi et al., 2021). For instance, non-random sampling of subgroups leads to sampling bias, which is comparable to representativeness bias (Shahbazi et al., 2022). Trends calculated for one population might not transfer to data gathered from a different population due to this bias (Mehrabi et al., 2021). One sort of selection bias is self-selection bias, in which study participants choose themselves. A poll that gauges support for a political candidate provides an example of this kind of bias, as those who are more excited about the candidate are more likely to fill out the survey (Siddique, 2023). Three types of bias in data may be distinguished: cognitive, selection, and reporting bias (Mikołajczyk-Bareła & Grochowski, 2023). When the actors, demographic information, statistics, and user attributes of a platform's user base diverge from the initial intended audience, demographic bias is another kind of prejudice that occurs (Schwartz et al., 2022). Furthermore, historical bias—a prejudice that already exists because of socio-technical problems in the world—can affect the data creation process even when the sample and feature selection are perfect (Mehrabi et al., 2021). Siddique (2023) mentions cognitive bias as a type of data bias. Different user behaviors across platforms, situations, or datasets might result in behavioral bias (Mikołajczyk-Bareła & Grochowski, 2023). Similarly, variations in populations and behaviors throughout time give rise to temporal bias (Schwartz et al., 2022). Additionally, there is content generation bias, which results from variations in user-generated material's structure, vocabulary, semantics, and syntax (Siddique, 2023). The last problem is sampling bias, which happens when systematic mistakes in data collecting result in a data sample used to train an algorithm that is not representative of the total population (Shahbazi et al., 2022).

3.2 Algorithmic Bias

Algorithmic bias relates to the workings of the models themselves (Mehrabi et al., 2021; Schwartz et al., 2022). Research has indicated that algorithmic bias has two subtypes: estimator bias and optimizer bias (Hastie et al., 2009; Vardi, 2023; Barrett & Dherin, 2021). Algorithmic bias also includes regularization bias (Barrett & Dherin, 2021; Vardi, 2023). In addition to data bias as a source of unfairness, there are biases that emerge during algorithm development (Suresh & Guttag, 2021). Evaluation bias also appears during model evaluation, including the use of inappropriate and inconsistent criteria for evaluating applications (Schwartz et al., 2022; Suresh & Guttag, 2021). When dealing with temporal data, we encounter the so-called longitudinal data fallacy (Mehrabi et al., 2021). Although cross-sectional analysis is frequently used to analyze temporal data, researchers should utilize longitudinal analysis to monitor groups over time to understand their behavior (Mehrabi et al., 2021). Among the bias categories associated with data consolidation, aggregation bias is also highlighted (Suresh & Guttag, 2021). When network attributes based on users' relationships, interactions, or activities differ, linkage bias occurs, which skews users' real behavior (Doidge & Harron, 2019; Li et al., 2022). Lastly, use and contact with actual users might lead to the emergence of prejudice. This bias usually appears after the design is finished and is caused by shifts in community knowledge, cultural beliefs, or demography (Schwartz et al., 2022).

3.3 User Bias

Individuals' actions while dealing with systems are closely linked to user prejudice (Jin et al., 2023). The user interface and the user themselves, through their self-selected biased behavior and interaction, are two causes of user interaction bias, for instance, which is a kind of bias that is not just seen online (Huang et al., 2024; Jin et al., 2023). One effect of the way information is presented is presentation bias. Internet users are only able to click on the material they view, for instance, therefore only the stuff they view gets clicks (Klimashevskaia et al., 2024). Ranking bias follows, in which more clicks are drawn to results that are rated higher because they are perceived as more significant and relevant (Tennenholtz et al., 2023). There is also popularity bias, whereby more popular items tend to appear more frequently, but popularity metrics are susceptible to manipulation—for example, through fake reviews or social bots (Klimashevskaia et al., 2024). Furthermore, social bias emerges when the actions of others influence our judgment (Jin et al., 2023). Individual bias cannot be ignored. Due to inherent human biases, biases are present in all available data (Jin et al., 2023; Huang et al., 2024).

3.4 Engineer Bias

This type of bias arises from the bias of the developer or engineer when designing the model. Implicit bias occurs when engineers make assumptions about the model they are training based on their personal experiences and beliefs, which may not necessarily apply to everyone (Ferrara, 2023). Confirmation bias also occurs when machine learning engineers process data without being aware of the data content and train models in a way that confirms their initial hypotheses or beliefs (Ferrara, 2023). Experimenter bias refers to evaluating or processing data in a way that aligns with the model developer's prior expectations or beliefs (Ferrara, 2023).

3.5 Social Bias

Social bias refers to bias that arises from a background social system and its accompanying social policies. The problems associated with models are more often in the data from which they were trained and not the algorithm itself. Models may apply to data that was generated from humans, or the human data generated as a secondary effect of social or historical inequities (Pagano et al., 2023; Ferrara, 2023).

4 . FAIRNESS AND ITS MEASUREMENT

Fairness in decision-making refers to the absence of bias or favoritism towards any person or group because of their characteristics, whether innate or acquired (Caton & Haas, 2020). By this reasoning, an unfair algorithm is one that makes biased decisions due to its membership in a group of people (Verma & Rubin, 2018). There is an appropriate construct for fairness is the use of equal probability, equal opportunity, and demographic equality. While these measures of fairness highlight the importance of a significant aspect of eliminating bias (Verma & Rubin, 2018; Alves et al., 2022). The concern of bias is related to an area of research called methodologies, referred to as fairness; fairness methodologies describe the importance of making neutral, non-biased machine learning models (Caton & Haas, 2020; Alves et al., 2022). If machine learning is mechanism to study decision making, then fairness is our ethical lens for analyzing this decision making process (Barocas, Hardt, & Narayanan, 2019). In other words, this means an act or decision was fair to the extent it discriminated against a person based on their status in a category (i.e., gender, race, etc.) (Verma & Rubin, 2018). Fairness metrics have recently emerged to measure the fairness of the decisions made by a model. These metrics take into account three parameters: A represents the characteristic to be measured, such as race or gender; S represents the predicted outcome; and Y represents the actual value.

4.1 Demographic Parity

This measure is used to ensure that the probability of a positive outcome is equal across groups, regardless of the sensitive characteristic and is expressed by the following equation:

$$P(S = 1 \mid A = 0) = P(S = 1 \mid A = 1) \quad \text{.....(1)}$$

The probability of a positive prediction should be the same across sensitive groups

4.2 Equal Opportunity

This measure is based on true positive outcomes, meaning that an individual who truly deserves a positive outcome (such as a patient in need of treatment or a student who is eligible to succeed) should be given an equal chance by the algorithm, regardless of which category they fall into. It is expressed as follows:

$$P(S = 1 \mid A = 0, Y = 1) = P(S = 1 \mid A = 1, Y = 1) \text{ .....(2)}$$

The True Positive Rate (TPR) must be equal across sensitive groups for individuals with $Y = 1$

4.3 Equalized Odds

It is a measure of fairness that measures the equality of the false positive rate (FPR) across groups. in other words, the model should neither overattribute errors nor underestimate the probability of any group occurring compared to any other group. It is expressed as the following equation:

$$P(S = 1 \mid A = 0, Y = y) = P(S = 1 \mid A = 1, Y = y), \forall y \in \{0,1\} \text{ .....(3)}$$

4.4 Disparate Impact

It measures the proportion of positive decisions between two groups. According to the "80% rule," the model is considered fair if the proportion of the less fortunate group is no less than 80% of the proportion of the other group.

$$\frac{P(S=1|A=0)}{P(S=1|A=1)} \geq 0.8 \text{.....(4)}$$

5. Ensemble Learning And Fairness

Ensemble learning refers to a set of machine-learning techniques that utilize two or more machine-learning models, or learners, and combine their outputs to create improved predictions (Gohar et al., 2023). Ensemble models combine two or more individual model outputs to make more accurate predictions than is possible from either of the members of an ensemble (Feffer et al., 2023). The goal of ensemble learning is to create better predictive performance than either of the individual models by using inputs from each of the individual models to make a group prediction that, when combined with inputs from other models, is a more accurate output than is possible from any one of the individual models. Ensemble learning in machine learning is classified into two categories: parallel learning methods and sequential learning methods (Grgić-Hlača et al., 2017). The most commonly used ensemble methods in machine learning are boosting, bagging, and stacking (Iosifidis & Ntoutsi, 2019). Many studies have been done on many different forms of ensemble learning methods and how they can be used in practice (Carrizosa et al., 2025).Ensemble learning and Fairness have a two-way relationship; first, ensemble method(s) can improve fairness by reducing prediction or prediction variability, and thus improving the chance of having predictions that do not include outliers caused by imbalanced training data. For example, Bagging minimizes random fluctuations in prediction and minimizes any random fluctuations that might occur to underrepresented or overrepresented classes (Iosifidis et al., 2019), and while Boosting can decrease the amount of bias in prediction, it will also amplify bias in

prediction if not experienced in the training data (Iosifidis & Ntoutsi, 2019), so unless specifically built around fairness-oriented strategies, boosting model(s) will only enhance discrimination or bias caused by data or bias itself.. Stacking and Voting offer flexibility by combining heterogeneous models, which may allow balancing between predictive accuracy and fairness (Ko et al., 2023).Research has explicitly explored integrating fairness objectives into ensemble methods. Fairness-aware ensembles either adjust data distributions (pre-processing), modify learning algorithms (in-processing), or recalibrate predictions (post-processing) (Chen et al., 2022; Zhang et al., 2023). This integration enables ensemble learning not only to boost accuracy but also to act as a bias mitigator (He et al., 2023).

5.1 Pre-processing approaches

In this approach, data are modified prior to the training phase, through reweighting, resampling, or selecting members that are consistent with fairness considerations. The goal is to reduce the inheritance in the data before it is transferred to the model. He et al. (2023) proposed BAHEM based on oversampling and post-processing... to mitigate data bias and improve model classification fairness with a slight decrease in prediction accuracy. Iosifidis et al. (2020) introduced the fair-bagging FAE method as a pre-processing mechanism to balance sensitive groups prior to model training, reducing discrimination and achieving high balanced accuracy. Conahan et al. (2025) used a pre-training reweighting method, EEC-GIFT, to improve the balance between Black and White subgroups, which, combined with thresholding, reduced racial bias in lung cancer screening. Mishra and Kumar (2023) iFselect pre-selects fair ensemble members, ensuring individual fairness without using sensitive attributes, resulting in a fairer ensemble.

Table 1. Pre-processing Approaches

Authors / Year	Algorithm	Fairness Metrics	Results
He et al. (2023)	Stacking Ensemble + Clustering-based Oversampling	AOD, EOD, ACC, BA	Oversampling alleviated bias; improved fairness with little accuracy reduction.
Iosifidis et al. (2020)	Bagging + Boosting with Fair-Bagging	Demographic Parity, Equalized Odds	Pre-processing balanced groups, reducing bias before training.
Conahan et al. (2025)	Ensemble + Reweighing (pre) + GIFT thresholding (post)	SPD, EOD	Reweighing improved subgroup balance prior to training.
Mishra & Kumar (2023)	Outlier Detection Ensembles with Individually fair member selection	iFscore (individual fairness)	Pre-selection ensured fairer ensemble composition without sensitive attributes.

5.2 In-processing approaches

In this approach, fairness is introduced during the learning process itself, by modifying the algorithm or adding multiple objectives to achieve a balance between accuracy and fairness. The goal is for the algorithm to become fairness-aware during model building, not just before or after training. Konak and Rodriguez (2025) proposed MELF, which uses the concept of ensemble learning with bagging to train an ensemble model that can improve the fairness of predictions without compromising accuracy. MELF relies on concepts from multi-objective decision making… to identify "good" models with respect to both predictive performance and fairness metrics. Chen et al. (2022) introduced MAAT, an ensemble framework that separately integrates fairness-optimized and accuracy-optimized models. This framework outperformed traditional methods in 92% of cases. AdaFair (2022) proposed a modified fairness-aware version of AdaBoost, where unfairly classified instances were weighted, thus reducing discrimination during boosting. Zhang et al. (2023) presented a multi-objective evolutionary framework, where "an evolutionary ensemble learning framework is introduced to mitigate unfairness while maintaining predictive accuracy, simultaneously improving collective and individual fairness." Lassig et al. (2022) introduced FALCES, proposing the use of dynamic selection strategies to achieve local and global fairness, ensuring fair treatment at different levels with limited impact on performance. Similarly, FAIR-Ensemble (2023) showed that the results confirm the intuition that ensembles often improve the stability of both accuracy and fairness measures… Minority groups improved disproportionately as group size increased." Feffer et al. (2023) conducted extensive experiments using eight bias mitigators from AIF360 with the methods of Bagging, Boosting, Voting, and Stacking… The results confirmed that ensembles often improve the stability of fairness measures. Finally, Carrizosa et al. (2025) presented an EFTE model where "a mixed linear optimization model is proposed to construct fair and interpretable tree clusters, enhancing fairness and interpretability simultaneously." While Grgić-Hlacha et al. (2017) presented "a theoretical analysis of fairness, diversity, and randomness in clustering methods, demonstrating the inherent trade-offs between fairness and accuracy."

Table 2. In-processing Approaches

Authors / Year	Algorithm	Fairness Metrics	Results
Konak & Rodriguez (2025)	Bagging + Multi-objective optimization (Pareto, TOPSIS)	SPD, EO, EOdds	Balanced fairness–accuracy trade-off; Pareto-based ensemble selection.
Chen et al. (2022)	Hybrid ensemble combining fairness-optimized + accuracy-optimized models	AUC, EOD	Outperformed SOTA in 92% of experiments.
AdaFair (2022)	Fairness-aware AdaBoost (upweights unfair samples)	SPD, AOD, EOD	Improved fairness during boosting by reweighting biased instances.
Zhang et al. (2023)	Multiobjective Evolutionary Stacking	Group fairness (fG), Individual fairness (fI)	Improved fairness across metrics not optimized during training.
Lässig et al. (2022)	Ensembles (dynamic selection)	Global & Local Equal Opportunity	Enhanced local fairness while maintaining global fairness.
Wei-Yin et al. (2023)	Deep homogeneous ensembles	Minority subgroup accuracy; disparate impact	Minority groups showed disproportionate accuracy gains with larger ensembles.
Feffer et al. (2023)	Comparison across Bagging, Boosting, Stacking, Voting + 8 AIF360 mitigators	DI, SPD, EOD, AOD	Ensembles improved stability; best setup dataset-dependent.
Carrizosa et al. (2025)	MILO-based Fair & Explainable Tree Ensembles	Accuracy across sensitive subgroups	Enhanced fairness and interpretability simultaneously.
Grgić-Hlača et al. (2017)	Theoretical analysis of fairness, diversity, randomness	Conceptual fairness metrics	Showed trade-offs between fairness, diversity, and accuracy.

5.3 Post-processing Approaches

These approaches rely on modifying the model outputs after the training phase, by recalibrating the decision boundary, adjusting the prediction probabilities, or making subsequent adjustments to the clustering results. Their primary goal is to improve fairness without the need to retrain the models from scratch, making them practical solutions for many applications. He et al. (2023) presented BAHEM, a hybrid model based on oversampling with a post-processing phase. This phase helped correct unfair outputs, resulting in improved fairness with a very slight decrease in prediction accuracy. IOSIFIDIS et al. (2020) proposed the FAE framework, which employs threshold-shifting methods to modify decision boundary placement, or classification thresholds, toward reducing overall bias while retaining a fair degree of accuracy. CONAHAN et al. (2025) employed the EEC-GIFT technique, a post-training method employing GIFT thresholding, to reduce bias associated with race-based disparities in predictions related to lung cancer screening over traditional training methods. CHEN et al. (2024) created the FairHOME approach which integrates input mutations and post-hoc voting. FairHOME was shown to improve intersectional equality by nearly 47% with virtually no decrease in performance measures relative to other methods. BHAGARVA et al. (2020) developed the LimeOut method, which is based on feature dropout ensembles, providing interpretations post-training with LIMEGlobal. LimeOut has a lower reliance on sensitive features than some other approaches, but has shown an accuracy of over 84% in practical application. Tran et al. (2021) created the PATE (Private Aggregation of Teacher Ensembles) technique, which aggregates the results from teacher models used to teach data to create privacy-based results based on statistical aggregation methods. PATE evaluates the effect of privacy on fairness as it computes fairness-related metrics, including Demographic Parity and Equal Opportunity, based on which type of aggregation method is selected for computation.

Table 3. Post-processing Approaches

Authors / Year	Algorithm	Fairness Metrics	Results
He et al. (2023)	Stacking Ensemble + CAN-based Post-processing	AOD, EOD	Post-processing corrected unfair outputs, further improving fairness.
Iosifidis et al. (2020)	Bagging + Boosting + Threshold shifting	Equal Opportunity	Adjusted decision thresholds to reduce bias.
Conahan et al. (2025)	Easy Ensemble + Reweighing + GIFT thresholding	SPD, EOD	Threshold adjustment mitigated racial bias while maintaining predictive power.
Chen et al. (2024)	Input mutation + Ensemble Voting/Averaging + Post-hoc adjustment	WC/AC-SPD, WC-AOD, WC-EOD	~47% improvement in intersectional fairness with minimal performance loss.
Bhargava et al. (2020)	Feature Dropout ensembles + Voting + LIMEGlobal (post-hoc)	Process Fairness, Accuracy	Reduced reliance on sensitive features; ~84% accuracy.
Tran et al. (2021)	Private Aggregation of Teacher Ensembles (post-aggregation)	Demographic Parity, Equal Opportunity	Explored fairness–privacy trade-offs; fairness affected by aggregation choices.

5.4 Analysis

The literature on bias and fairness in ensemble learning algorithms has presented a clear difference in the intervention approaches undertaken. Some research relied on preprocessing approaches that modified the data before any training, for example, reweighting or oversampling. The results have shown that these approaches are easy to implement and are capable of improving balance between groups, but often have been inadequate to eliminate bias due to the structural mechanisms of the algorithm. Other research has used in-process approaches that used fairness constraints within the learning process itself, using multi-objective approaches or fairness-sensitive boosting. This body of work has shown that this normative category is capable of achieving a clearer balance between accuracy and fairness but has been more complex and demanding of computational resources. Finally, the research using post-processing approaches has provided flexibility in which fairness can be improved by resetting thresholds or changing outputs after training, without the need to reconstruct the model entirely. Some experiments have even shown the ability to improve fairness metrics, with only a negligible loss in accuracy.

Looking at the differences in methods shows that preprocessing is a good place to start for remediating biases in data, though its effectiveness can be minimal if the bias arises from the learning process. Preprocessing during training offers a more complete approach, but with higher costs in complexity and time, and postprocessing is a workable approach to correct outputs while getting the benefit of reasonable performance. The literature has also shown a variation in how fairness metrics are defined. Some authors have focused on collective fairness metrics (such as demographic parity and equal opportunity), while others have opted for more complex metrics (such as individual fairness and intersectional fairness). This inconsistency is due to the difficulty of achieving consensus on a single metric to determine fairness, and the need for an analysis framework to evaluate all of the different metrics as a whole to assess models comprehensively. In spite of successful work done to date, the literature indicates that there is still no clear guidance on how to optimally balance accuracy and fairness; improving fairness often comes at the expense of decreasing predictive power, and vice versa. Additionally, ensemble-learning models lack a clear means of interpretation, making it difficult to know why a model may be biased. Therefore, it is imperative to develop new and refined frameworks to evaluate fairness, accuracy, and interpretability, and apply these methodologies in areas where errors could potentially have severe impact, such as medical and financial sectors and those within criminal justice.

6. CHALLENGES AND FUTURE DIRECTIONS

A number of key challenges facing the enhancement of Fairness in Ensemble learning Algorithms, which can be summarised as follows: A balance between Accuracy and Fairness, "Ensuring Fairness in an Ensemble Model is often in exchange for Predictive Accuracy, and so achieving Equally Balanced Trade-off remains one of the most important Challenges to date."Interpretability is limited, In particular, Stacking and Boosting methods decrease interpretability, thus making it challenging to identify the sources of bias and justify the decision-making process.

Multiplicity of Metrics and Standardisation is complex; due to the many Different types of Fairness Metrics, including Group Fairness (e.g., demographic parity) as well as Individual and Intersectional Fairness, Comparative analyses across studies are made more difficult and prevent the creation of a Standard Evaluation Framework. Application in Sensitive Domains; the investigation of Fairness-aware Ensemble Models has primarily been undertaken using Standard Benchmark Datasets (e.g., COMPAS and Adult Data); Investigating the Application of Fairness in Ensemble Models in Higher Stakes Domains such as Healthcare, Criminal Justice and Financial Services has yet to occur. Inherent Bias exists in Data, The continuation of Data-level bias (i.e., historical bias, and demographic bias) through the Ensemble Models, despite pre-processing activities, indicates that further, deeper solutions will be necessary to resolve this Issue. Computational Resources and Complexity: In-Process Fairness-aware Methods (especially those based on multi-objective optimization) are often computationally expensive to implement, limiting their ability to Scale as Practical Implementations in the Real World.

7.FUTURE DIRECTIONS

Balancing Fairness and Explainability: The future will focus on developing even better ensemble methods for understanding, trusting and embracing fairness through viable transparency. Fairness-Oriented Deep Learning Ensembles: This is an exciting avenue to explore, especially as we have seen a rise in the number of entities who are using deep learning in the more critical areas. Metrics Development and Intersectionality: In order to move away from the way fairness is identified by groups, we must create metrics for assessing fairness with intersecting identities and context Framework Development Around Multi-Objective Optimization Characteristics: The development of frameworks for multi-objective optimization providing the ability to maximize the performance metrics of fairness, interpretability, and accuracy has a high likelihood of becoming an emerging area of future research interest. Field Testing Fairness-Oriented Ensemble Techniques: In order to close the significant gap that exists between the currently available theoretical modelling constructs of fairness and the actual implementation of

that technology within the industry, empirical evidence relating to using fairness-oriented ensemble methods across multiple industries must be generated immediately.

8.Conclusion

The review has shown that many ensemble algorithms have the potential to improve the quality and stability of a prediction but still fail when there is bias in the dataset or algorithmic design that produces bias. A number of different solutions have been proposed in the literature to solve this issue, including modifying the data prior to training, adding fairness constraints during training, or adjusting model output after it has been produced. In terms of the comparative review of these solutions, it is apparent that some of these solutions are better than others at balancing both accuracy and fairness; however, there is still a need to develop a more unified way of measuring fairness in all of these situations. Finally, given the significant advances in the area of fairness and bias in ensemble algorithms, there is an additional need to develop an integrated framework that combines both fairness and interpretability, as well as efficiency, for use in multiple critical fields (e.g., healthcare, law enforcement, finance, etc.). As a result, this paper aims to provide an overview of the current research that addresses bias and fairness within ensemble learning algorithms, so that future research efforts can effectively develop fairer and more reliable systems.

REFERENCES

[1] Alves, G., Bernier, F., Couceiro, M., Makhlouf, K., Palamidessi, C., & Zhioua, S. (2022). Survey on fairness notions and related tensions. arXiv preprint arXiv:2209.13012. https://doi.org/10.48550/arXiv.2209.13012

[2] Arévalo-Cordovilla, F. E., Cárdenas-Estévez, D., & Saénz-de-Sicilia, V. (2025). Evaluating ensemble models for fair and interpretable early prediction of academic performance. Scientific Reports, 15, 2234. https://doi.org/10.1038/s41598-025-15388-9

[3] Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning: Limitations and opportunities. MIT Press. https://fairmlbook.org

[4] Barrett, D. G. T., & Dherin, B. (2021). Implicit gradient regularization. In International Conference on Learning Representations (ICLR 2021). https://openreview.net/forum?id=3q5IqUrkcF

[5] Bhargava, V., Couceiro, M., & Napoli, A. (2020). LimeOut: An ensemble approach to improve process fairness. arXiv preprint arXiv:2006.10531. https://doi.org/10.48550/arXiv.2006.10531

[6] Carrizosa, E., Kurishchenko, K., & Romero Morales, D. (2025). On enhancing the explainability and fairness of tree ensembles. European Journal of Operational Research, 323(2), 599–608. https://doi.org/10.1016/j.ejor.2025.01.008

[7] Caton, S., & Haas, C. (2024). Fairness in machine learning: A survey. ACM Computing Surveys, 56(4), 1–40. https://doi.org/10.1145/3616865

[8] Chen, Z., Zhang, J. M., Sarro, F., & Harman, M. (2022). MAAT: A novel ensemble approach to addressing fairness and performance bugs for machine learning software. In Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '22) (pp. 1122–1134). ACM. https://doi.org/10.1145/3540250.3549093

[9] Chen, Z., Li, X., Zhang, J. M., Sarro, F., & Liu, Y. (2024). Diversity drives fairness: Ensemble of higher order mutants for intersectional fairness of machine learning software. arXiv preprint arXiv:2412.08167. https://doi.org/10.48550/arXiv.2412.08167

[10] Conahan, P., Robinson, L. A., Le, T., Valdes, G., Schabath, M. B., Byrne, M. M., Green, L., El Naqa, I., & Luo, Y. (2025). Easy ensemble classifier-group and intersectional fairness and threshold (EEC-GIFT): A fairness-aware machine learning framework for lung cancer screening eligibility using real-world data. JNCI Cancer Spectrum, 9(2), pkaf030. https://doi.org/10.1093/jncics/pkaf030

[11] Doidge, J. C., & Harron, K. L. (2019). Reflections on modern methods: Linkage error bias. International Journal of Epidemiology, 48(6), 2050–2060. https://doi.org/10.1093/ije/dyz203

[12] Feffer, M., Hirzel, M., Hoffman, S. C., Kate, K., Ram, P., & Shinnar, A. (2023). Searching for fairer machine learning ensembles. In Proceedings of the 2nd International Conference on Automated Machine Learning (AutoML '23). https://doi.org/10.48550/arXiv.2210.05594

[13] Ferrara, E. (2025). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. AI, 6(1), 3. https://doi.org/10.3390/ai6010003

[14] Ford, J. (2025). Quantitative insights into bias within machine learning risk evaluations and established credit models. Financial Innovation, 11(1), 43. https://doi.org/10.1186/s40854-025-00543-4

[15] Gohar, U., Biswas, S., & Rajan, H. (2023). Towards understanding fairness and its composition in ensemble machine learning. In Proceedings of the 45th International Conference on Software Engineering (ICSE '23) (pp. 1533–1545). IEEE. https://doi.org/10.1109/ICSE48619.2023.00137

[16] Hasanzadeh, F., Rahimi, M., & Luo, J. (2025). Bias recognition and mitigation strategies in artificial intelligence in healthcare. npj Digital Medicine, 8, 46. https://doi.org/10.1038/s41746-025-01503-7

[17] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning (2nd ed.). Springer.

[18] He, F., Wu, X., Zhang, W., & Huang, X. (2023). A novel bias-alleviated hybrid ensemble model based on over-sampling and post-processing for fair classification. Connection Science, 35(1), 2184310. https://doi.org/10.1080/09540091.2023.2184310

[19] Hu, L., Wang, Y., Zhang, R., & Sun, J. (2024). Enhancing fairness in AI-enabled medical systems with the attribute neutral framework. Nature Communications, 15, 52930. https://doi.org/10.1038/s41467-024-52930-1

[20] Huang, Y., Lyu, S., & Wu, X. (2024). A scoping review of fair machine learning techniques when using real-world data. Journal of Biomedical Informatics, 150, 104640. https://doi.org/10.1016/j.jbi.2024.104640

[21] Iosifidis, V., & Ntoutsi, E. (2019). AdaFair: Cumulative fairness adaptive boosting. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19) (pp. 781–790). ACM. https://doi.org/10.1145/3357384.3358141

[22] Iosifidis, V., Fetahu, B., & Ntoutsi, E. (2019). FAE: A fairness-aware ensemble framework. In Proceedings of the IEEE International Conference on Big Data (BigData '19) (pp. 1375–1380). IEEE. https://doi.org/10.1109/BigData47090.2019.9005647

[23] Jin, D., Wang, L., Zhang, H., Zheng, Y., Ding, W., Xia, F., & Pan, S. (2023). A survey on fairness-aware recommender systems. Information Fusion, 100, 101906. https://doi.org/10.1016/j.inffus.2023.101906

[24] Ko, W. Y., D'souza, D., Nguyen, K., Balestriero, R., & Hooker, S. (2023). FAIR-Ensemble: When fairness naturally emerges from deep ensembling. arXiv preprint arXiv:2303.00586. https://doi.org/10.48550/arXiv.2303.00586
[25] Lässig, N., Oppold, S., & Herschel, M. (2022). Metrics and algorithms for locally fair and accurate classifications using ensembles. Datenbank-Spektrum, 22(1), 23–43. https://doi.org/10.1007/s13222-021-00401-y
[26] Lindloff, C., & Siegert, I. (2025). Defining bias in AI-systems: Biased models are fair models. arXiv preprint arXiv:2502.18060. https://arxiv.org/abs/2502.18060
[27] Mishra, G., & Kumar, R. (2023). An individual fairness based outlier detection ensemble. Pattern Recognition Letters, 171, 1–7. https://doi.org/10.1016/j.patrec.2023.05.010
[28] Ni, H., Han, L., Chen, T., & Zhang, C. (2024). Fairness without sensitive attributes via knowledge sharing. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24) (pp. 123–133). ACM. https://doi.org/10.1145/3630106.3658904
[29] Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Peixoto, R. M., Guimarães, G. A. S., Cruz, G. O. R., Araujo, M. M., Santos, L. L., Cruz, M. A. S., & Oliveira, E. L. S. (2023). Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. Big Data and Cognitive Computing, 7(1), 15. https://doi.org/10.3390/bdcc7010015
[30] Pham, N., Le, T., & Tran, D. (2025). FAIREDU: A multiple regression-based method for improving fairness across multiple sensitive features. Expert Systems with Applications, 245, 123086. https://doi.org/10.1016/j.eswa.2024.123086
[31] Raftopoulos, G., Ntoutsi, E., & Iosifidis, V. (2025). A comprehensive review and benchmarking of fairness-aware decision trees and ensemble methods. Algorithms, 18(7), 435. https://doi.org/10.3390/a18070435
[32] Raftopoulos, G., Papamitsiou, Z., & Giannakos, M. (2025). Evaluating fairness strategies in educational data mining. Electronics, 14(9), 1856. https://doi.org/10.3390/electronics14091856
[33] Schwartz, R., Vassilev, A., Greene, K. K., Perine, L., Burt, A., & Hall, P. (2022). Towards a standard for identifying and managing bias in artificial intelligence (NIST Special Publication 1270). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.1270
[34] Siddique, S. (2023). Survey on machine learning biases and mitigation. Computers, 12(1), 1. https://doi.org/10.3390/computers12010001
[35] Singh, A., & Ghosh, J. (2025). Bias mitigation through proxy sensitive attribute labeling. In Proceedings of the 2025 ACM Conference on AI, Ethics, and Society (AIES '25) (pp. 210–220). ACM. https://doi.org/10.1145/3703323.3703329
[36] Tennenholtz, G., Mladenov, M., Merlis, N., Axtell, R. L., & Boutilier, C. (2023). Ranking with popularity bias: User welfare under self-amplification dynamics. arXiv preprint arXiv:2305.18333. https://arxiv.org/abs/2305.18333
[37] Tran, C., & Fioretto, F. (2023). On the fairness impacts of private ensemble models. arXiv preprint arXiv:2305.11807. https://doi.org/10.48550/arXiv.2305.11807
[38] Verma, S., & Rubin, J. (2018). Fairness definitions explained. In Proceedings of the IEEE/ACM International Workshop on Software Fairness (FairWare '18) (pp. 1–7). ACM. https://doi.org/10.1145/3194770.3194776
[39] Vardi, G. (2023). On the implicit bias in deep-learning algorithms. Communications of the ACM, 66(5), 46–55. https://cacm.acm.org/research/on-the-implicit-bias-in-deep%E2%80%90learning-algorithms/
[40] Yang, Y., Wang, C., & Li, H. (2024). A survey of recent methods for addressing AI fairness and bias. Journal of Biomedical Informatics, 147, 104621. https://doi.org/10.1016/j.jbi.2024.104621
[41] Yang, J., Soltan, A. A. S., et al. (2023). Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning. Nature Machine Intelligence, 5(7), 652–664. https://doi.org/10.1038/s42256-023-00697-3
[42] Zhang, Q., Liu, J., Zhang, Z., Wen, J., Mao, B., & Yao, X. (2023). Mitigating unfairness via evolutionary multiobjective ensemble learning. IEEE Transactions on Evolutionary Computation, 27(4), 848–862. https://doi.org/10.1109/TEVC.2022.3209544
[43] Zhang, Q., Wen, J., & Yao, X. (2024). Multi-objective ensemble machine learning for fairness. International Journal of Computational Intelligence Systems, 17(2), 233–247. https://doi.org/10.1007/s44196-024-00088-1

BIOGRAPHIES OF AUTHORS

Author 1 picture	Dr. Omar Shakir , Received His Phd. degree in the Doctor of computer science / Artificial intelligence from University of mosul . He has a teacher in the directorate of education in Nineveh
2 Author 2 picture	Dr. Maan younis , Received His Phd. degree in the Doctor of computer science / Artificial intelligence from University of mosul . He has a teacher in the directorate of education in Nineveh