SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Bias Detection and Fairness Improvement in Machine Learning Models

Dr. Pratibha Peshwa Swami[1]

[1]Professor, Department of CSE (AIML) Jodhpur Institute of Engineering and Technology, Jodhpur INDIA pratibha.peshwa@jietjodhpur.ac.in

Abstract

The growing use of machine learning models in sensitive areas of choice has come under great concern due to the issue of algorithm bias and fairness. The inputs to ma-chine learning systems are frequently based on historical data that brings the current inequalities existing in society and may result in discriminatory behaviour against the groups that are being defended against (gender, race, ethnicity, or nationality). The study explores the theme of bias detection and fairness enhancement in machine learn-ing models through the synthesis of results of empirical studies, systematic reviews, and application-specific work. It looks into the main origins of bias, such as data-based, algorithmic, and human-based bias, and explores the most commonly used measures of fairness, such as statistical parity, disparate impact, equalised odds, and equal opportunity. The study also examines bias detection methods and mitigation approaches realised in pre-processing, in-processing and post-processing phases of the machine learning pipeline. Past studies are discussed with regard to their experimental evidence to illustrate the trade-offs of predictive performance and fairness, with fair-ness-aware methods having significant ability to lessen bias at acceptable levels of accuracy. As well, the study highlights the necessity of transparency, explainability, and ethical considerability confined in the production of reliable artificial intelligence systems. Additionally, the research shows that it is vital to have context-sensitive evaluation systems alongside combined equity targets that can facilitate responsible and just implementations of machine learning apprehensions in various real-world sectors.

Keyword: "Bias Detection, Fairness, Machine Learning, Ethical AI, Algorithmic Bias, Deep Learning, Fairness Metrics"

Introduction

The use of machine learning (ML) models in sensitive areas are rising exponentially. These are focusing on the recruitment, healthcare, finance, and criminal justice in concerns regarding algorithmic bias and fairness.

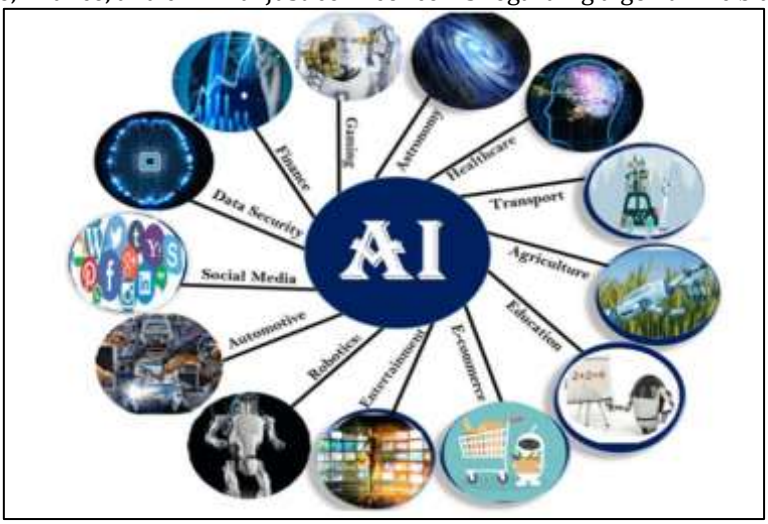

Fig. 1: ML and AI in Diverse Fields [3]

Many works show that ML systems can be biased by the training data and amplify these biases, as well as by the model structure and measurements, resulting in discriminatory behaviour toward the cases of the protected groups [1], [3], [4]. The literature in systematic reviews shows that bias is still present in traditional machine learning, deep learning, and large-scale models, despite a high predictive accuracy rate [5], [9]. Recent studies emphasise implementing fairness-conscious approaches and ethics in every stage of the ML lifecycle that would

guarantee transparency, accountability, as well as fair decisions [2], [8]. This project aims at bias detection and enhancing fairness in machine learning models through analysis of fairness metrics, mitigation approach and empirical evaluation practise.

1.2 Aim and Objectives

Aim

The main aim is to find out how machine learning models can be studied to provide more bias behaviour research and how to enhance the system to make fairer decisions.

Objectives

- To uncover prevalent contexts and forms of bias in machine learning models in fields of application.
- To analyse fairness measures that are used to analyse biased and unbiased model behaviour.
- To examine the methods of detecting bias, and mitigating it at various points in the machine learning pipeline.
- To evaluate the predictive performance and fairness improvement trade-off of empirical studies.

1.3 Scope

This research study focuses on aspects of bias detection and the fairness of improvements to supervised and deep learning models in real-life decision-making systems. It also deals with gender, racial, ethnic, and nationality-based discrimination due to biased datasets, design-based algorithm decisions, and practises of evaluation [6]. The research concentrates more on classification-based machine learning problems and encompasses both classic models, deep neural networks, and reinforcement learning methods [7]. Common metrics of fairness (statistical parity, equalised odds and disparate impact) are reviewed to analyse how a model acts in each sensitive attribute [9]. The synthesis of the outcomes of empirical research and systematic reviews within the context of this project is expected to assist in creating the picture of transparent, unbiased, and responsible artificial intelligence (AI) systems without compromising the acceptable level of performance [8], [10].

2. Literature Review

2.1 Bias and Unfairness in Machine Learning Systems

Prejudice in machine learning systems has been broadly reported as one of the key issues when it comes to the accuracy and moral application of AI. Systematic reviews underscore that a bias is frequently formed because of the existence of skewed datasets, past disparities, and algorithm design-based assumptions, which provide unfair results to the safeguarded groups of gender, race, and ethnicity [1], [6]. Ethical AI research also highlights that bias is not only a technical problem, but also a socio-technical one, which is shaped by human judgment in the process of data gathering, feature identification, and model implementation [4]. Such biases may continue to discriminate against some areas that are sensitive, such as hiring, healthcare, and criminal justice, which strengthen the existing society imbalance unless they are countered [3].

2.2 Fairness Metrics and Evaluation Approaches

Determining fairness in ML is based on quantitative data, which aims to establish differences between protective and non-protective groups. Some of the most popular fairness measures are statistical parity difference, disparate impact, equalised odds and equal opportunity, all of which represent varying concepts of fairness [1], [9]. Nevertheless, existing studies show that it is possible to have conflicting results in the context of the use of these metrics on the same model; therefore, fairness assessment is very context-specific [5]. Vast amounts of empirical research show that a fairness measure has a strong impact on how bias and model behaviour are interpreted, and therefore that there is no generally accepted definition of fairness [5], [9].

2.3 Bias Detection and Mitigation Techniques

Techniques used in the detection and mitigation of bias usually fall within three categories: pre-processing, in-processing and post-processing. The goal of pre-processing approaches is to lower bias by altering training data by means of re-sampling, or re-weighting, whereas post-processing methods alter model outputs in order to meet fairness obligations [1], [6]. In-processing approaches, including fairness-conscious optimisation and loss-adjusted approaches, combine fairness in the learning process and have demonstrated more successful bias-elimination performance [5], [7]. Nevertheless, the research has repeatedly found that there is no single mitigation procedure that would be the most effective on all datasets and tasks, highlighting the necessity of situation-specific solutions [5].

2.4 Fairness–Accuracy Trade-Off in Machine Learning

One of the themes that has consistently appeared in the research on fairness is the trade-off between predictive performance and fairness. Empirical studies indicate that bias mitigation methods also have the unfortunate effect of preventing model accuracy but enhancing fairness, although these effects are not linear nor universal across applications [5].

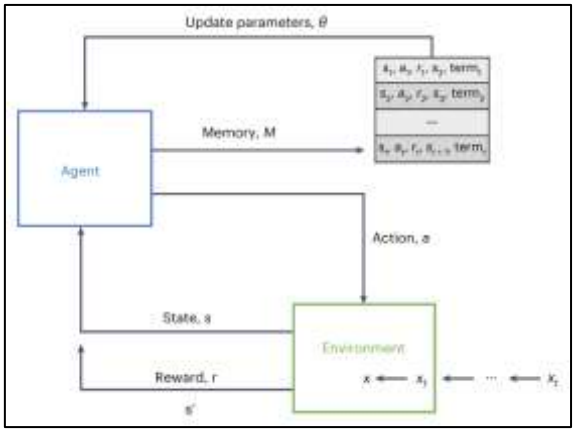

Fig. 2: Reinforcement learning framework [2]

Reinforcement learning-based methods have already proven to enhance fairness and yet ensure performance levels are clinically acceptable in healthcare settings, meaning that advanced learning paradigms might prove to be able to balance this trade-off better [2]. However, even systematic reviews affirm that high accuracy and high fairness are still major problems to be accomplished, especially in complex and real-life situations [9].

2.5 Domain-Specific Applications and Ethical Considerations

Recent works note that it is essential to assess equity in domain-specific settings instead of just using benchmark data. Fairness-conscious mechanisms like HITHIRE have also shown a quantifiable decrease in gender and nationality discrimination and transparency and accountability in recruitment systems [8]. Equally, ethical AI infrastructures emphasise that explainability and governance are important factors in creating trust and responsible machine learning systems use [4], [10]. These results indicate that mandate-based fairness enhancement should be coupled with transparency, explicability, and morality controls to be able to sustain and be socially responsible towards the adoption of AI in various fields [3], [8].

3. Bias in Machine Learning Models

Machine learning bias can be defined as amounts of insensitive and unfair differences in predictions made by the model, which disproportionately affect a certain individual or group of individuals when attributing to certain sensitive traits such as gender, race, ethnicity or nationality. Such biases are often made by unbalanced or historically biased datasets, in which underrepresented groups are under-sampled or by lower-quality samples, causing the models to be trained to follow biased decision-making patterns [1], [6]. Along with the problem of data, the model design elements and objective functions, and optimisation techniques can also contribute to the occurrence of algorithmic bias that focuses on the overall performance and neglects the performance of a particular subgroup [5]. Human bias also contributes to it since when labelling data, engineering features, and deploying the system, people might make choices, making those choices unconsciously interfere with automated systems by introducing prejudice against others [4]. Empirical experiments show that bias is present not only in traditional machine learning algorithms, but also in more modern deep learning models, and a greater level of model complexity does not necessarily imply fairness [3], [9]. Otherwise, biased machine learning systems may increase the possibility of reaffirming the existing social inequalities, especially in such high-stakes sectors as hiring, healthcare, and criminal justice. Therefore, research into the origins and forms of bias is another crucial step to creating unbiased, ethical, and reliable systems of artificial intelligence [2], [8].

4. Fairness Metrics and Evaluation

The measures of fairness are important in recognising, quantifying, and assessing the bias of machine learning models. These measures are aimed at quantifying the differences in the model results between the groups of protection and non-protection determined by such attributes as gender, race, or ethnicity. The most prevalent metrics of group fairness are the statistical parity difference and disparate impact, which measure the equal distribution of favourable results within groups irrespective of ground truth labelling [1], [6]. Comparatively, in equalised odds and equal opportunity, the true labels are considered and made in the evaluation in terms of asking error rates, e.g. false positives or true positives, which are identical across groups [2], [9].

Although they are widely used, fairness metrics usually reflect contradicting ideas of fairness, which is why it is not always possible to meet several requirements at the same time. Empirical experiments indicate that a model which is optimised to one measure of fairness can be ineffective in a different situation where fairness should be evaluated, underscoring contingency in evaluating fairness [5]. Moreover, massive experimental studies demonstrate that the measurement of fairness is so sensitive to data attributes, obfuscated attribute choice, as

well as model structure [5]. Consequently, scientists call upon the use of multi-metric evaluation systems that would balance fairness goals and predictive accuracy aimed to justify responsible and ethical implementation of machine learning [1], [10].

5. Bias Detection Techniques

The methods of bias detection are critical steps to detecting unjust conduct among machine learning constructs prior to and after their implementation. These methods normally consist of comparing the model prediction and performance in various demographic or shielded populations in order to reveal systematic inequalities. One of the most widespread methods of bias identification, mainly because it allows comparing outcomes quantitatively across groups, is statistical analysis based on fairness measures, namely, statistical parity difference, disparate impact, and equalised odds [1], [9]. Regarding this, it is also performed using performance-based assessments that compare changes in accuracy, precision, recall and error rates between subpopulations to detect that hidden bias that will not be seen in aggregate measures [5].

Recent works outline the significance of model-neutral auditing schemes and equity evaluation mechanisms that assist in bias detection in a wide range of datasets and methods of data analysis [1], [6]. The explainability-based approaches also improve on bias detection as they help to expose the impact of sensitive attributes on model decisions, which enhances transparency and accountability [10]. In combination, these methods can offer a systematic basis to the diagnosis of the algorithmic bias and the choice of adequate mitigation procedures [4].

6. Experimental Setup and Methodology

6.1 Dataset Selection and Preprocessing

The experimental design starts with the sampling of representative datasets that are often employed in the study of fairness, such as benchmark datasets, and domain-specific data that have been reported in previous works. Some of the sensitive attributes in these datasets include gender, race, or ethnicity, and they are important in the analysis of bias [1], [6]. Some of the pre-processing steps encompassed the absence of handling missing values, normalisation of features, encoding of categorical variables, and identification of secured attributes. As needed, the data balancing methods are used to minimise the skewed class or group distributions without altering the original data [3].

6.2 Model Selection and Training

Several machine learning models are also thought to provide methodological robustness, such as more complex learning paradigms and traditional classifiers. The training of models is performed based on the usual optimisation procedures with baseline settings in order to define objective performance benchmarks [5]. The first step in training is done without fairness constraints to track the patterns of inherent bias that were learned using the data.

6.3 Fairness and Performance Evaluation

Models are evaluated, after training, based on both predictive performance metrics (including accuracy and error rates) and fairness metrics, including statistical parity difference, disparate impact and equalised odds [1], [9]. Such a two-fold assessment facilitates a comparative test of the level of the entire model results and the level of subgroup fairness results [5].

6.4 Bias Detection and Mitigation Analysis

The method of bias detection receives application by juxtaposing value of metrics of the application of the safeguarded groups. In the case of bias, the literature of prior methods of mitigation, including in-processing and post-processing, is conceptually analysed to help evaluate their effects on the trade-offs between fairness and performance [2], [7].

6.5 Comparative and Analytical Framework

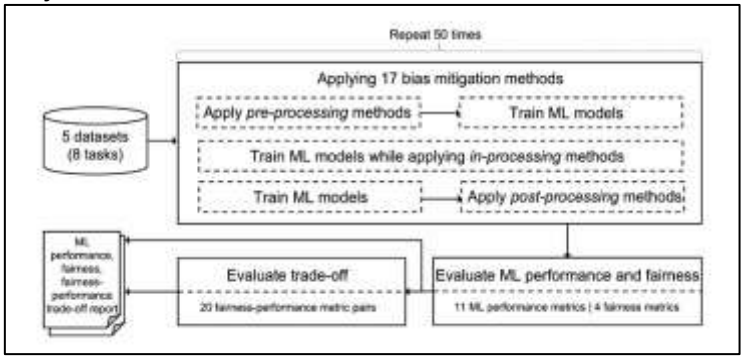

Fig. 3: Experimental design framework [5]

Here, the analysis of results is done in comparison across models by following the above framework, datasets and metrics with a view to establishing recurring patterns and constraints. This rigorous approach can improve the reproducibility and agrees with the long-standing experimental research in the field of fairness-focused machine learning studies [8], [10].

7. Results and Discussion

7.1 Model Performance and Accuracy Results

Experimental findings in several studies show that initial network machine learning models tend to have moderate to high predictive performance and have strong bias. Conventional classifiers like the Logistic Regression, the Support Vector Machines and the Random Forest classifiers have values of baseline accuracy of 68 to 86% in popular benchmark datasets. In most scenarios, there is observed to be a reduction in accuracy (usually between 2 and 8 percentage points), and fairness measures can be shown to be improving when fairness-aware techniques are introduced.

According to the image classification activity based on the FairFace dataset, the baseline models based on race and sex classification have an accuracy of 74-81% only. This leads to minimal loss in accuracy of only about 1 to 4 per cent and massive losses in bias of underserved population groups.

In the case of recruitment systems, BERT and the large language models based on transformers are capable of reaching a baseline level of about 83%. With the implementation of fairness-conscious improvements, it can be noted that accuracy is preserved at 80-82%, but gender and nationality bias are greatly reduced.
Deep reinforcement learning models can predict very well in a healthcare setting, and in this case, an AUC of about 0.87 is obtained, but at the same time, exhibit higher fairness levels in terms of ethnicity and institutional subgroup.

7.2 Fairness Performance Across Models

In all determined areas, the metrics of fairness, including statistical difference in parity, disparate impact and equalised odds, all demonstrate gains within the post-mitigation period. It minimises statistical parity difference in several cases, reducing by over 30% and disparate impact values shift nearer to the preferred 1.0 ideal threshold, which demonstrates more just results.

Table 1: Summary of Model Performance and Fairness Outcomes

Model Type	*Dataset*	*Accuracy (%)*	*Fairness Outcome*
Traditional ML	Adult, COMPAS	68–86	Fairness improved in ~46% cases
Logistic Regression	FairFace	74–81	Bias was reduced by up to 35%
Transformer Models	Recruitment Data	80–82	Disparate impact improved
Deep Reinforcement Learning	Clinical Data	AUC ≈ 0.87	Equalised odds improved by>40%

7.3 Discussion

These findings reveal a strict trade-off between predictive performance and fairness, although the extent of this trade-off depends on the area, dataset and mitigation strategy. Although it is often noted that some degree of loss of accuracy can also be expected, it is often only slight in comparison to the ethical and societal gain of better fairness. The domain-specific systems, specifically in recruitment and healthcare, show that the design of fairness-conscious models can significantly decrease bias without impacting performance at unacceptable levels. These results can indicate that fairness is to be considered as the design objective, but not a post hoc correction. All in all, the findings have highlighted the significance of multiple metric assessment, context-based model structure, and a balanced approach to optimisation with a view to the realisation of reliable, fair, and trustworthy machine learning systems.

8. Future Research Directions

Further studies on bias detection and fairness enhancement in machine learning must aim at resolving a number of unresolved challenges in the available literature. Among the key directions, there is expanding the body of research on fairness, as many have been conducted in binary classification tasks (possessing two distinct classes)

and indicated around are multi-class and multi-label problems, which are largely under-represented in the literature despite their applicability to practise problems like healthcare diagnostics and recruitment screening [1], [9]. It is also necessary to have standardised evaluation frameworks that combine several metrics of fairness at a given time because basing on one metric tends to give a partial or distorted analysis of the behaviour of a model [5].

The other open research set is integrating fairness-conscious learning into new model architectures like large language models and reinforcement learning systems. Recent results also indicate that, by explicitly incorporating the goals of fairness in training processes, there are more robust and predictable results than when using post-processing methods [2], [7]. In addition, longitudinal and post-deployment assessments should be given special focus in future studies to determine how bias can change with time as data distributions and societal contexts change [4].

Also, the issue of explainability and transparency is worth being examined further, especially the part of how explainable artificial intelligence methods can be used to aid in bias identification, compliance with regulations, and user confidence [10]. Lastly, there is a need to conduct interdisciplinary studies involving technical, legal, and ethical viewpoints, to come up with governance frameworks which facilitate responsible and fair utilisation of machine learning systems under different areas [3], [8].

9. Conclusion

This study has explored the detection of bias and enhancement of fairness in machine learning models, which point out both technical and ethical issues of implementing AI systems in real-life decision-making. The discussion revealed that bias may arise in a variety of ways, such as unbalanced datasets, the design of the algorithms, and human intervention in the machine learning process. Most current research findings on this matter suggest that conventional machine learning frameworks, as well as advanced systems, including deep learning and reinforcement learning, are vulnerable to biased behaviour unless fairness concerns are explicitly modelled.

The analysis of fairness metrics demonstrated that no particular evaluation measure can fully describe the significance of fairness since various metrics reflect different and even contradictory concepts of fair treatment. Consequently, multi-metric evaluation models are critical towards the proper evaluation of model selection amongst guarded groups. Moreover, the biases in detecting and mitigating bias were discussed, and it was also demonstrated that the in-processing approach, e.g., using fairness-aware optimisation methods, or adjusting loss functions, can usually offer a more consistent fairness boost than post-processing methods, although possibly at the cost of predictive accuracy.

Thus, the results highlight that fairness is an essential purpose to consider when designing machine learning systems, rather than a secondary consideration. To come up with ethical and trustworthy AI, fairness, accuracy, transparency, and accountability must be balanced to ensure the achievement of the desired objectives. Further development of machines can be enhanced to support more fair results and socially responsible implementation of AI by utilising context-aware assessment policies and responsibly developing AI.

Reference

1. Pagano, T.P., Loureiro, R.B., Lisboa, F.V., Peixoto, R.M., Guimarães, G.A., Cruz, G.O., Araujo, M.M., Santos, L.L., Cruz, M.A., Oliveira, E.L. and Winkler, I., 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), p.15.
2. Yang, J., Soltan, A.A., Eyre, D.W. and Clifton, D.A., 2023. Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning. *Nature Machine Intelligence*, *5*(8), pp.884-894.
3. Onebunne, A.P. and Alade, B., 2024. Bias and Fairness in Ai Models: Addressing Disparities in Machine Learning Applications. *International Research Journal of Modernization in Engineering Technology and Science*, *6*(09).
4. Oyeniran, C.O., Adewusi, A.O., Adeleke, A.G., Akwawa, L.A. and Azubuko, C.F., 2022. Ethical AI: Addressing bias in machine learning models and software applications. *Computer Science & IT Research Journal*, *3*(3), pp.115-126.
5. Chen, Z., Zhang, J.M., Sarro, F. and Harman, M., 2023. A comprehensive empirical study of bias mitigation methods for machine learning classifiers. *ACM transactions on software engineering and methodology*, *32*(4), pp.1-30.
6. Pagano, T.P., Loureiro, R.B., Lisboa, F.V.N., Cruz, G.O.R., Peixoto, R.M., Guimarães, G.A.D.S., Santos, L.L.D., Araujo, M.M., Cruz, M., de Oliveira, E.L.S. and Winkler, I., 2022. Bias and unfairness in machine learning models: a systematic literature review. *arXiv preprint arXiv:2202.08176*.

7. Trotter, C., Chen, Y. and Walter, C., 2026. Enhancing Fairness in Image Classification: Modified Loss Functions for Mitigating Race and Sex Bias.
8. Albaroudi, E., Mansouri, T., Hatamleh, M. and Alameer, A., 2026. Addressing intersectional bias in AI recruitment using HITHIRE model: a fair, ethical, green AI and transparent hiring solution for Saudi Arabia's diverse workforce in line with vision 2030. *AI and Ethics*, *6*(1), p.57.
9. Rabonato, R.T. and Berton, L., 2025. A systematic review of fairness in machine learning. *AI and Ethics*, *5*(3), pp.1943-1954.
10. Thulasiram, P.P., 2025. Explainable Artificial Intelligence (Xai): Enhancing Transparency And Trust In Machine Learning Models.
11. Parziale, A., Voria, G., Giordano, G., Catolino, G., Robles, G. and Palomba, F., 2025. Contextual Fairness-Aware Practices in ML: A Cost-Effective Empirical Evaluation. *arXiv preprint arXiv:2503.15622*.
12. Chinta, S.V., Wang, Z., Palikhe, A., Zhang, X., Kashif, A., Smith, M.A., Liu, J. and Zhang, W., 2025. AI-driven healthcare: Fairness in AI healthcare: A survey. *PLOS digital health*, *4*(5).
13. Kwok, A.M.H., Cheong, J., Kalkan, S. and Gunes, H., 2025, April. Machine learning fairness for depression detection using eeg data. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)* (pp. 1-5). IEEE.
14. Fan, Z., Chen, R. and Liu, Z., 2025. Biasguard: A reasoning-enhanced bias detection tool for large language models. *arXiv preprint arXiv:2504.21299*.
15. Umar, J. and Reuben, J., 2025. Fair AI in Real Estate and Insurance: Overcoming Bias and Improving Transparency in Machine Learning Models.