Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 5 Issue 1 (2025)

Bias Detection and Mitigation in HR Algorithms: Ensuring Fairness in Machine Learning Models

Neha Jain[1]
Assistant Professor,
Computer Science and Engineering, Poornima University, Jaipur
neha.jain@poornima.edu.in
Prateek Agrawal[2]
Sr. Solution Architect,
Masters of Computer Applications, Gurukula Kangri (Deemed to be University)
ag.prateekg@gmail.com
Dr. G. Kumar[3]
Assistant Professor,
Faculty of Management, SRM Institute of Science and Technology,
Kattankulathur, Chennai, Tamilnadu
kumarg@srmist.edu.in
Dr. T. Velmurugan[4]
Assistant Professor, Faculty of Management,
SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamilnadu
velmurut@srmist.edu.in
Dr. R. Naveenkumar[5]
Associate Professor,
Department of Computer Science and Engineering,
Chandigarh Colleges of Engineering,
Chandigarh Group of Colleges, Jhanjeri, Mohali -140307, Punjab, India
rnaveenkumarooty@gmail.com
Ms. Ramandeep Kaur[6]
Assistant Professor,
Department of Management, Chandigarh School of Business,
Chandigarh group of Colleges Jhanjeri, Mohali, Punjab – 140307

Abstract:-
Additional efforts are necessary to guarantee that the machine learning algorithms employed in healthcare do not perpetuate or exacerbate any preexisting discriminatory or objectionable biases that may be present in the data. In order to reduce the impact of any biases that may have developed during the data collection procedure, we implement a reinforcement learning approach. We assessed the effectiveness of our model in accurately predicting the incidence of COVID-19 in patients seeking medical attention at hospital emergency departments in order to reduce the impact of potential biases associated with ethnicity and hospital location. By utilizing an innovative incentive function and training strategy, we have proven that our methodology outperforms the most advanced machine learning techniques in terms of clinically significant screening outcomes and equitable results. To illustrate the adaptability of our methodology, we implemented a comprehensive evaluation of our approach in an intensive care unit by administering a discharge status test. Additionally, we conducted external validation at three distinct institutions.

Keywords:- Medical ethics, translational research, and diagnosis

I. INTRODUCTION

It is impossible to exaggerate the importance of fairness in machine learning algorithms, particularly in light of the growing dependence of organisations on these systems for HR tasks. Algorithms are frequently employed to make objective decisions, including those concerning career advancement, performance evaluations, and hiring. Nevertheless, machine learning algorithms have the potential to produce biassed results by inadvertently perpetuating the biases that are already present in the training data. In order to guarantee that all individuals are treated fairly, it is essential to address prejudice in HR algorithms, as this issue has significant ethical, legal, and societal repercussions.

Bias in HR algorithms can be caused by a variety of factors, such as human error, inadequate model design, and biassed training data. A model may unintentionally continue to exhibit a preference for certain demographics over others if it is trained using recruiting data that reflects societal preconceptions. This could result in a lack of diversity, unjust recruitment procedures, and unequal opportunities. Firms that strive to foster inclusive work environments are therefore morally obligated to identify and rectify bias in these algorithms, in addition to being a technical challenge.

These concerns have been addressed by scientists and professionals who have developed a variety of methods to identify and reduce bias in machine learning algorithms. The objective of these strategies is to guarantee equity by minimising bias during data preparation, preventing discrimination during the learning process, and addressing any unfairness in post-processing outcomes [1]. In order to implement these strategies effectively, it is imperative to have a comprehensive understanding of the specific context in which the HR algorithms are implemented. Additionally, it is imperative to consistently monitor and assess the algorithms to ensure that there is no bias.

The integration of machine learning into HR operations has the potential to significantly improve the efficacy and decision-making processes. Nevertheless, it also enhances the probability of individuals' preexisting beliefs being reinforced, which can have an effect on both groups and individuals. Firms should prioritise the identification and mitigation of bias in HR algorithms to guarantee fairness and equality in their HR procedures. This will allow them to fully utilise the capabilities of machine learning.

II. RELATED WORKS

The increasing prevalence of machine learning models in talent management and recruitment has given rise to a growing need for research on the detection and mitigation of algorithm bias in HR. Numerous studies have demonstrated that HR algorithms are prejudicial, particularly in relation to age, gender, and race. This bias has the capacity to lead to discriminatory recruitment and promotion decisions [2]. The primary objective of the initial research on this subject was to identify the underlying factors that contribute to bias in ML models. Barocas and Selbst (2016) illustrated that discriminatory behaviours in AI systems may result from biassed training data and model design decisions, thereby perpetuating historical societal disparities.

Further research has been conducted to investigate methods for detecting and quantifying prejudice in HR algorithms as a result of these discoveries. This issue has been resolved by the emergence of fairness-aware machine learning algorithms. In order to train models that effectively mitigate bias, these methods implement constraints. Feldman et al. (2015) have developed algorithms that alter decision constraints to facilitate the comparison of results across various demographic groups [3]. Alternative methodologies, including those proposed by Hardt et al. (2016), implemented concepts such as equalised odds and differential impact to establish impartiality metrics that could be implemented consistently during model training.

In order to eradicate bias in HR algorithms, numerous remedial measures have been implemented. The solutions that have been investigated include a variety of pre-processing methods to reduce bias, in-processing techniques to modify the learning algorithm, and post-processing methods to modify the output of biassed models. Recent research (e.g., Kamiran and Calders 2012) indicates that the inclusion of fairness requirements in model design can effectively mitigate discriminatory results without compromising the overall performance of the model [4].

Additionally, adversarial debiasing has become increasingly popular due to its ability to train models that are resistant to bias. Zhang, Lemoine, and Mitchell (2018) illustrated the utilisation of adversarial networks to penalise predictions that are biassed. These developments emphasise the ongoing endeavours to develop HR algorithms that are both highly effective and adhere to ethical standards of justice and equity [5].

The necessity of continuous research and innovation in the identification and mitigation of algorithmic bias in HR has been underscored by previous studies. Despite the increasing use of machine learning models in HR operations, ensuring their impartiality remains a significant challenge that necessitates a comprehensive approach that takes into account advancements in algorithmic design, data processing, and ethical considerations [6].

III. RESEARCH METHODOLOY

Our primary focus is on clinical applications when addressing bias and equity, which is a result of three fundamental reasons. Will begin by addressing the current issue: inaccurate predictions made by a biassed model may be the basis for significant decisions that could have a significant impact on one's life. Additionally, biases against specific demographic groups may contribute to discrepancies in the quality of healthcare provided to specific patient cohorts [7]. Additionally, a biassed model has the potential to exacerbate existing healthcare and societal disparities. The deployment of ML models in clinical practice is impeded by the presence of these challenges, which diminish the level of confidence between physicians and their patients.

a) Data Collection and Preprocessing

The discipline of medicine is plagued by racial discrimination. Unintentional biases in data collection, such as those that influence the selection of individuals for admissions, volunteers, samples, or observers, may result in conclusions that do not accurately represent the entire population.

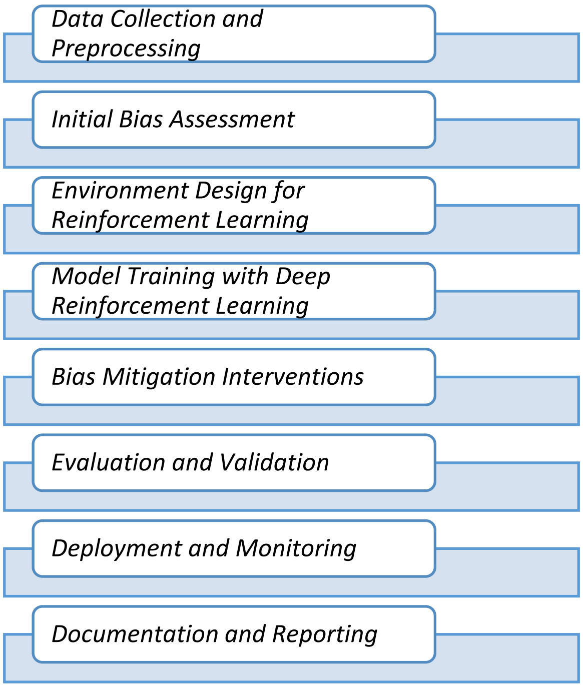
Data Collection and Preprocessing
Initial Bias Assessment
Environment Design for Reinforcement Learning
Model Training with Deep Reinforcement Learning
Bias Mitigation Interventions
Evaluation and Validation
Deployment and Monitoring
Documentation and Reporting

Fig.1: Flow diagram for the proposed methodology.

Previous research has demonstrated that ethnicity-related biases may be present in machine learning algorithms. The recidivism prediction algorithm exhibited racial bias by disproportionately misclassifying Black offenders as potential felons in comparison to white defendants. This error resulted in a substantial increase in the probability of such misclassifications, which was approximately twice as high [8]. The efficacy of machine learning models was inconsistent when applied to various patient populations in clinical settings, as discovered by researchers. This could potentially have a detrimental effect on under-represented groups.

b) Initial Bias Assessment

This is a significant issue, as the research sample populations frequently do not accurately represent the entire patient population as a result of factors such as financial constraints and regional biases. The demographic profile of the actual patient population that will receive the medication may not be consistently reflected in the clinical trial participants. Despite the utilisation of randomised trials to assess treatment outcomes in the trial population this disparity continues to exist [9]. Consequently, the most basic level of healthcare remedies may be provided to women, individuals from non-white ethnic backgrounds, and those with a higher body mass index if decisions are made based on models. In situations where there is a restricted number of patients from a specific ethnic group, the biassed machine learning model increases the probability of identifying individuals. The protection of confidential information and the dissemination of statistical data are both impacted by these implications.

c) Environment Design for Reinforcement Learning

Disparities in the quality of healthcare services, the occurrence of diseases, mortality rates, and the incorporation of specific medical developments can result from variation in the care provided by hospitals and clinics. The age range is 14 to 18 years old. Measurement bias is the unintentional inclusion of biases that are unique to a specific location during the collection, analysis, and organisation of data. Consequently, machine learning models that are exclusively based on data from a single hospital may not be generalisable to multiple contexts [10]. An example of this is a research that discovered that sophisticated computer vision techniques were frequently used to misdiagnose underprivileged patient groups.

d) Model Training with Deep Reinforcement Learning

Many clinical projects that employ machine learning aspire to consolidate datasets from multiple institutions in order to augment the quantity of training data, as generalisability frequently necessitates larger sample sizes. This mitigates the probability of fallacies. As a consequence of variations in the quantity of training data that are accessible, models may develop biases that are unique to each training centre [11]. These biases have the potential to influence a model's decision-making and result in subpar performance for specific healthcare facilities by exacerbating disparities among hospitals and restricting the implementation of machine learning technologies.

e) Bias Mitigation Interventions

In order to address these concerns, practitioners are increasingly adopting methods to address the issues of impartiality and bias reduction in machine learning. These solutions can be implemented on three distinct levels: algorithms, data, and evaluation. Our primary goal is to employ the RL paradigm to develop a fair model at the algorithmic level [12]. The utilisation of standard supervised learning in conjunction with adversarial debiasing is the most prevalent method in contemporary research for mitigating algorithmic bias. A model is trained to learn parameters that do not disclose sensitive features as part of this approach. The adversarial network ensures that the prediction network's output is not influenced by the sensitive feature provided during the training process, which is the bias we are attempting to eliminate.

f) Evaluation and Validation

Supplementary Section B also specifies that a fairness metric may function as a constraint or be incorporated into a loss function. This method has been implemented to create precise forecasting models that effectively mitigate gender disparity in pay prediction (with a sample of 6 males and 20 females) and racial prejudice in recidivism prediction (with individuals of black and white ethnicity) 21. Moreover, adversarial models have been employed to enhance the fairness of results by diminishing ethnic and location-based biases, as well as to provide precise predictions for COVID-19 [13]. To improve the impartiality of the results, we elected to employ a reinforcement learning (RL) framework rather than an adversarial one. This decision is made due to the necessity of specific strategies to mitigate unfavourable biases.

g) Deployment and Monitoring

Reinforcement learning (RL) is an artificial intelligence technique that involves an agent learning a task through interaction with its environment. RL has been associated with numerous significant achievements in the field of real-world AI. This includes notable occurrences in the fields of gaming and control. On the other hand, the fundamental concepts of RL have been shown to be effective in a broader range of activities, including those that do not appear to involve a single "agent" interacting with a "environment" (which is the typical scenario for RL) [14]. Traditional supervised learning techniques are frequently employed to resolve these concerns, particularly in classification assignments. Reinforcement Learning (RL) employs an agent to evaluate the input, classify it, and subsequently receive an instantaneous reward from its environment in accordance with the prediction.

h) Documentation and Reporting

The agent receives a positive reward when it accurately predicts the label, and a negative reward when it fails to do so. The agent can determine its optimal "behaviour" for accurately detecting samples using this data, thereby achieving the highest possible reward. In order to accomplish this, agents participate in activities that produce memory cells. Agents employ these fields in conjunction with the initial input to classify samples and select actions [15]. Prior research has shown that the implementation of specific reward functions can effectively address significant data imbalances associated with the desired label. We have developed a sophisticated deep reinforcement learning (RL) framework that is designed to enhance algorithmic fairness and mitigate biases (as illustrated in Figure 1), rather than solely concentrating on the correction of label imbalance.

IV. RESULTS AND DISCUSSION

This research introduced a deep reinforcement learning framework that was employed to train unbiased machine learning models. We subjected it to a rigorous evaluation on two complex real-world tasks—predicting patient discharge status and COVID-19 screening—while simultaneously striving to mitigate biases associated with the hospital environment and patient ethnicity.

Our research demonstrated that RL's robust classification performance was not compromised, and the impartiality of outcomes was significantly improved. Comparing RL to existing benchmarks and state-of-the-art ML techniques, such as XGBoost, RL (without debiasing), and a fully connected neural network (NN) with or without cost-adjusted weights inversely proportional to the frequency of sensitive attributes, and with or without adversarial debiasing, this improvement was observed.

Removing Ethnic Prejudice

Our models were trained using patient cohorts from three distinct NHS foundation trusts: University Hospitals Birmingham (UHB), Bedfordshire Hospitals NHS Foundation Trust, and Oxford University Hospitals (OUH). Subsequently, we implemented these models to verify our discoveries. Our models attained exceptional area under the receiver operator characteristic curve (AUROC) scores in all test sets, which is consistent with the results of previous studies3,19,26,27 that employed comparable patient groups and characteristics (see Supplementary Table 13).

Please refer to Table 1 for further details. This suggests that we initially trained classifiers that are robust and resilient. The BH cohort demonstrated the most impressive performance among the test groups, with an AUROC score that ranged from 0.897 to 0.923 (95% confidence interval: 0.861-0.954). The AUROC values for predicting COVID-19 status in the UHB and BH cohorts, on the other hand, were consistently maintained at 0.821 to 0.894, 0.834 to 0.868, and 0.807 to 0.892, respectively.
The traditional supervised learning models, including adversarial, weighted and unweighted XGBoost, and NN models, consistently obtained the highest AUROCs, despite the fact that all models produced similar AUROCs. The models achieved average AUROCs of 0.869 (RL), 0.857 (RL, unweighted), 0.881 (adversarial), 0.885 (NN), 0.881 (XGBoost), 0.887 (NN, weighted), and 0.879 (XGBoost, weighted).

When employing a sensitivity setting of 0.9, our analysis demonstrated that all models and cohorts exhibited consistent sensitivity scores. In the case of PUH, UHB, and BH, the sensitivity ranges were 0.862-0.879, 0.825-0.913, and 0.862-0.935, respectively. RL achieved the highest sensitivity ratings on both the UHB and BH test sets.

In addition, our models have generated substantial prevalence-dependent negative predictive value (NPV) scores (>0.978), suggesting that COVID-19 can be confidently excluded, as has been previously reported in other empirical investigations. Our findings suggest that the reinforcement learning (RL) paradigm is more suitable for diverse contexts, as it exhibits enhanced sensitivity towards the two most ethnically distinct cohorts, BH and UHB.. Furthermore, the RL paradigm demonstrates a higher AUROC (Area Under the Receiver Operating Characteristic curve) for the BH cohort.

Table.1: Equalized odds evaluation for ethnicity bias and COVID-19 status prediction test results across different models and test sets, optimized to sensitivities of 0.9

Test set	Model	EO (TP)	EO (FP)	F1
PUH				
	RL	**0.047**[a]	0.037	0.159
	RL (unweighted)	**0.048**[b]	0.031	0.160
	ADV	0.050	**0.014**[a]	0.186
	NN	0.066	**0.028**[b]	0.202
	XGB	0.133	0.053	0.172
	NN (weighted)	0.056	0.035	**0.210**
	XGB (weighted)	0.214	0.054	0.191

We employed threshold adjustment to guarantee a high level of sensitivity in our classification tasks, particularly in the context of predicting COVID-19 and ICU discharge. This strategy was advantageous as a result of the conflicting training data for both objectives. However, the data was biassed in the t-SNE visualisation as a result of the unique characteristics of each location (Fig.2). Consequently, the optimal threshold that was derived from a single dataset may no longer be appropriate when new conditions with differing distributions are encountered. This is a plausible explanation for the sensitivity disparities observed between test sites in the COVID-19 task that involves ethnicity debiasing.

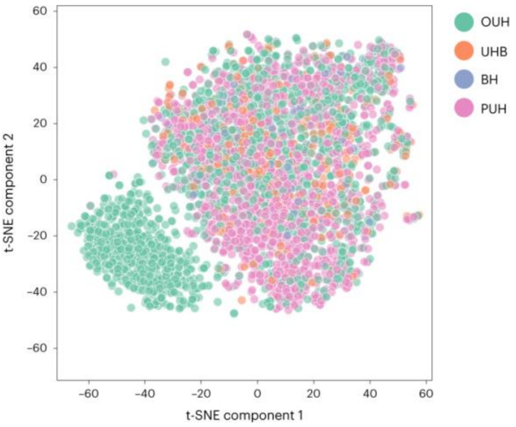

Fig.2: Denotes t-SNE representation.

Visualization of clusters determined through a t-SNE, including all positive COVID-19 cases across four NHS trusts (OUH, PUH, UHB, BH).

V. CONCLUSION AND FUTURE DIRECTION

Identifying and preventing bias in HR algorithms is essential for ensuring that the use of machine learning models for employee recruiting and management is fair and equitable. In order to promote a diverse and equitable workplace, organisations can identify and eliminate biases, thereby facilitating more impartial and inclusive decision-making. By conducting data audits, undertaking regular model reviews, and implementing fairness-aware machine learning techniques, HR procedures can be improved to significantly reduce the likelihood of biassed outcomes, thereby improving accuracy and fairness.

The primary objective should be to improve the efficacy of bias detection tools and to seamlessly integrate them into HR operations in the future. The collaboration of AI researchers, ethicists, and HR specialists is essential for the development of robust models. This collaboration is essential to guarantee that the models are not only technically sound but also comply with ethical standards. In order to guarantee equity as machine learning models develop, it will be imperative to consistently analyse and integrate a variety of datasets. Additionally, it is essential to prioritise transparency in algorithmic decision-making and continue to conduct research on innovative mitigation techniques to guarantee enduring impartiality in HR algorithms.

REFERENCES

[1]. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., and Bengio, Y. (2014). Generative Adversarial Networks. In Advances in Neural Information Processing Systems (pp. 2672–2680).

[2]. N. Yu, L. S. Davis, and M. Fritz (2019). Fake Image Attribution Using GAN Fingerprints: Learning and Analyzing. In the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 7556-7566.

[3]. Marra, F., D. Gragnaniello, D. Cozzolino, and L. Verdoliva (2018). Detection of GAN-generated Fake Images on Social Networks. In the IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 384–389).

[4]. Nataraj, L., She, H., Flenner, A., and Manjunath, B.S. (2019). Detecting GAN-generated false photos with co-occurrence matrices. Electronic Imaging, 2019(5): 532-1. Chai, H., Zhu, J., and Yin, X. (2021). GAN-Based Image Forgery Detection: A Survey. IEEE Access, 9, 49037–49057.

[5]. Kwon H., Yang H., Lim J., & Kim Y. M. (2020). Unsupervised Learning for GAN Image Forgery Detection. In Proceedings of the 28th ACM International Conference on Multimedia (pp. 1563–1571).

[6]. "Agnihotri, A. and Kapoor, S., Implications of Super Leadership and Self Leadership for Production Processes in Indian IT Sector, International Journal of Mechanical and Production Engineering Research and Development (IJMPERD) ISSN (P): 2249-6890; ISSN (E): 2249-8001 Vol. 8, Issue 3, Jun 2018, 875-886"

[7]. S. RadhaMahendran, A. Dogra, D. Mendhe, S. B. G. Tilak Babu, S. Dixit and S. P. Singh, "Machine Learning for Drug Discovery: Predicting Drug-Protein Binding Affinities using Graph Convolutional Networks," 2024 5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST), Jamshedpur, India, 2024, pp. 87-92, doi: 10.1109/ICRTCST61793.2024.10578506.

[8]. B. S. Panigrahi, B. Pattanaik, O. Pattanaik, S. B. G. Tilak Babu, P. G and B. Shaik, "Reinforcement Learning for Dynamic Power Management in Embedded Systems," 2024 5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST), Jamshedpur, India, 2024, pp. 51-55, doi: 10.1109/ICRTCST61793.2024.10578373

[9]. "Agnihotri, A., & Maurya, A. (2023). The Pandemic Benefits Reaped by Online Teaching Platforms: A Case study of Whitehat Junior. Journal of Information Technology Management, 15(3), 69-84. doi: 10.22059/jitm.2023.93625https://jitm.ut.ac.ir/issue_11529_12096.html https://doaj.org/article/d08f1fdc5e1044dfa7eb3a4479e9b048

[10]. "DeGroat, W., Abdelhalim, H., Patel, K. et al. Discovering biomarkers associated and predicting cardiovascular disease with high accuracy using a novel nexus of machine learning techniques for precision medicine. Sci Rep 14, 1 (2024). https://doi.org/10.1038/s41598-023-50600-8

[11]. A. Shahi, G. Bajaj, R. GolharSathawane, D. Mendhe and A. Dogra, "Integrating Robot-Assisted Surgery and AI for Improved Healthcare Outcomes," 2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 2024, pp. 1-5, doi: 10.1109/ICONSTEM60960.2024.10568646.

[12]. P.S. Ranjit, et al., Experimental Investigations on Hydrogen Supplemented Pinus Sylvestris Oil-based Diesel Engine for Performance Enhancement and Reduction in Emissions, FME Transactions, Vol 50. No.2 , pp. 313-321, 2022, doi:10.5937/fme2201313R

[13]. Ahmed Z, Zeeshan S, Mendhe D, Dong X. Human gene and disease associations for clinical-genomics and precision medicine research. Clin Transl Med. 2020; 10: 297–318. https://doi.org/10.1002/ctm2.28

[14]. P.S. Ranjit, Pankaj Sharma and Mukesh Saxena, “Experimental Investigations on influence of Gaseous Hydrogen (GH2) Supplementation in In-Direct Injection (IDI) Compression Ignition Engine fuelled with Pre-Heated Straight Vegetable Oil (PHSVO)” International Journal of Scientific & Engineering Research (IJSER), Volume 5, Issue 10, October 2014, ISSN: 2229-5518.

[15]. S. R. Mahendran, A. Dogra, D. Mendhe, S. B. G. T. Babu, S. Dixit and S. P. Singh, "Machine Learning-Assisted Protein Structure Prediction: An AI Approach for Biochemical Insights," 2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 2024, pp. 1-6, doi: 10.1109/ICONSTEM60960.2024.10568895.