Bias Detection in Clinically Machine Learning Models

MANJU V A[1], HARISH T A[2]
[1,2] *Channabasaveshwara Institute of Technology. Tumkur, Karnataka.*

Abstract-- ***The integration of machine learning (ML) into healthcare has enabled significant advances in diagnosis, prognosis, and treatment planning. However, predictive models often inherit biases from imbalanced datasets and structural inequities, leading to unfair outcomes across demographic groups. This project presents a Healthcare Bias Detection Tool, developed in Python with a Tkinter-based interface, to evaluate both the performance and fairness of clinical ML models. The tool supports algorithms such as Support Vector Machines (SVM) and K-Nearest Neighbors (KNN), while enabling users to load datasets, preprocess data, and configure sensitive attributes and target variables. Beyond traditional metrics like accuracy, precision, recall, and F1-score, the system incorporates fairness indicators including disparate impact ratio and statistical parity difference. Visualizations such as confusion matrices, subgroup accuracy comparisons, and prediction distributions enhance interpretability. The tool also generates structured reports covering performance, bias findings, and recommended mitigation strategies, thus fostering accountability and ethical AI adoption in healthcare. By combining model validation with fairness evaluation, the framework contributes to responsible deployment of machine learning in clinical decision-making.***

Index Terms- ***Machine learning, bias detection, fairness metrics, healthcare AI, clinical decision support, support vector machine (SVM), K-nearest neighbors (KNN), disparate impact, statistical parity difference, ethical AI, model evaluation, transparency in AI.***

I. INTRODUCTION

The integration of machine learning (ML) and artificial intelligence (AI) into healthcare has transformed the way clinical decisions are made, enabling advancements in disease diagnosis, patient risk stratification, treatment prediction, and personalized care. Predictive models have the potential to improve efficiency and patient outcomes; however, they are not immune to bias. When trained on imbalanced datasets or influenced by systemic inequities, these models may produce unfair or inaccurate results, disproportionately affecting vulnerable groups such as women, elderly patients, or underrepresented communities. Such disparities can lead to misdiagnoses, unequal treatment recommendations, and ultimately widen existing healthcare inequalities.

Bias in healthcare ML models can originate from multiple sources, including skewed training data, inappropriate feature selection, or inherent structural biases in electronic health records (EHRs). These hidden prejudices can result in significant consequences for clinical practice, thereby undermining trust in AI-driven healthcare systems. Ensuring fairness and transparency has thus become an essential requirement for the ethical deployment of ML in medicine. Regulatory authorities, hospitals, and researchers are increasingly emphasizing explainability, fairness-aware algorithms, and accountability frameworks to safeguard patient rights.

In response to these challenges, this project introduces a Healthcare Bias Detection Tool—an interactive Python-based application with a graphical user interface (GUI) built using Tkinter. The tool allows users to upload clinical datasets, configure predictive models, and evaluate model performance alongside fairness metrics across sensitive attributes such as gender, age, race, or socio-economic status. By integrating algorithms such as Support Vector Machines (SVM) and K-Nearest Neighbors (KNN), along with fairness indicators like disparate impact ratio and statistical parity difference, the tool provides a comprehensive environment for bias detection and performance evaluation.

Furthermore, the system enhances interpretability through subgroup-based accuracy analysis, confusion matrices, and visualization dashboards, ensuring that hidden disparities are revealed beyond overall accuracy. Structured reports are generated to summarize model performance, fairness assessment, and potential mitigation strategies such as data balancing, fairness constraints, or post-processing adjustments. This framework supports researchers, clinicians, and policymakers in fostering ethical AI adoption, promoting equity in healthcare delivery,

and ensuring that predictive models remain both reliable and fair in real-world applications.

A. Problem Statement

Machine learning models in healthcare often suffer from bias due to imbalanced datasets, underrepresentation of certain demographic groups, or limitations in algorithm design. Biased models may produce unequal predictions across sensitive groups such as age, gender, or ethnicity, leading to unfair treatment recommendations and reinforcing existing healthcare disparitie

B. Related work

The increasing use of machine learning in healthcare has created opportunities for disease diagnosis, prognosis, and personalized treatment. However, several studies have highlighted the risks of algorithmic bias and the urgent need for fairness-aware approaches.

Rajkomar et al. [1] emphasized that ML models can enhance healthcare outcomes but may also inherit biases from imbalanced datasets and underrepresentation of certain groups. Similarly, Obermeyer et al. [2] demonstrated how widely used health risk algorithms underestimated illness severity among Black patients due to biased cost-based proxies. Chen et al. [3] reported that classifiers often show discriminatory behavior, with performance differing across demographic subgroups, highlighting the clinical risks of biased predictions.

To address such challenges, researchers have proposed multiple fairness-aware strategies. Pre-processing techniques, such as re-weighting and re-sampling, were introduced by Kamiran and Calders [4] to reduce discrimination before training. Zafar et al. [5] explored fairness constraints integrated during the training phase, while Hardt et al. [6] proposed equality of opportunity as an in-processing method. Post-processing approaches, such as adjusting decision thresholds, have also been studied to reduce group-level disparities without retraining.

Recent reviews have highlighted healthcare-specific implications. Chen, Liu, and Lin [7] analyzed bias in electronic health records (EHRs), emphasizing how missing data and socio-economic proxies distort predictions. Cross, Mehta, and Singh [8] investigated fairness concerns in medical AI decision-making, stressing the ethical and trust-related implications of biased models. Huang, Zhang, and Li [9] provided a scoping review of fairness techniques, noting trade-offs between accuracy and fairness in healthcare applications. Poulain et al. [10] further explored federated learning as a potential approach to improve fairness across institutions while preserving privacy.

These studies collectively underline that fairness in ML is both an algorithmic and socio-ethical challenge. While existing works propose fairness metrics and mitigation methods, most tools remain technically complex and inaccessible to clinicians. The proposed Healthcare Bias Detection Tool builds on this body of research by offering an interactive, user-friendly framework that integrates model evaluation with fairness analysis, bridging the gap between technical fairness solutions and practical healthcare applications.

II. COMPARISON WITH PREVIOUS WORK

Previous studies in healthcare machine learning have primarily focused on predictive performance, with fairness considerations often addressed only as a secondary concern. Existing models and frameworks emphasize metrics such as accuracy, sensitivity, specificity, and AUC, but they rarely evaluate subgroup disparities in depth. As a result, models that perform well overall may still underperform for certain demographic groups, such as women, elderly patients, or racial minorities, leading to hidden inequities in clinical outcomes.

Some fairness-aware methods have been introduced in the literature. Pre-processing approaches such as re-weighting and re-sampling [4] aim to mitigate bias before training, while in-processing techniques like fairness constraints [5] and adversarial debiasing [12] embed fairness during model training. Post-processing strategies [6], on the other hand, adjust outputs after prediction to reduce disparities. While effective in controlled experiments, these methods often require significant technical expertise and are not easily accessible to clinicians or healthcare administrators.

Moreover, most existing tools lack integration of performance and fairness metrics within a single user-friendly framework. Clinicians and policymakers frequently rely on technical audits conducted by data scientists, limiting transparency

and interpretability for non-technical stakeholders. Reviews such as those by Chen et al. [7] and Cross et al. [8] highlight that current solutions are fragmented and not tailored for healthcare-specific workflows.

The proposed Healthcare Bias Detection Tool advances beyond prior work in several ways:

1. Unified Evaluation – It integrates both conventional performance measures (accuracy, precision, recall, F1-score) and fairness indicators (statistical parity difference, disparate impact ratio) within the same framework.

2. Accessibility – With a Tkinter-based graphical interface, it lowers the technical barrier, making bias detection usable by healthcare professionals without coding expertise.

3. Visualization and Interpretability – Unlike many prior works that rely solely on numerical outputs, the tool provides subgroup comparisons, confusion matrices, and prediction distribution plots for clearer interpretation.

4. Actionable Reporting – The system generates structured reports that summarize model behavior, highlight fairness concerns, and suggest mitigation strategies, bridging the gap between research prototypes and practical decision support.

By combining bias detection with interpretability and usability, this work addresses the limitations of previous approaches and provides a comprehensive framework for ensuring fairness in healthcare predictive modeling.

III. METHODOLOGY

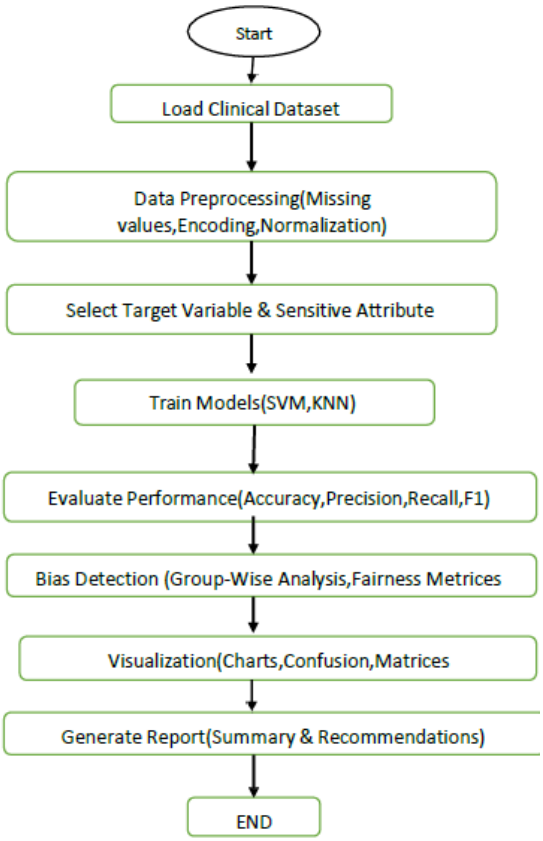

Fig 1: Methodology Flowchart

The proposed Healthcare Bias Detection Tool follows a structured methodology that integrates machine learning model development, fairness evaluation, and interpretability within a single framework. The workflow consists of five main stages:

1. Data Collection and Preprocessing

Clinical datasets are collected from electronic health records (EHRs), medical repositories, or open datasets. Preprocessing steps include handling missing values, outlier detection, normalization/standardization, encoding categorical features, and addressing class imbalance using techniques such as SMOTE. This ensures data quality and fairness readiness before model training.

2. Model Training

The system allows users to select classification algorithms such as Support Vector Machines (SVM) and K-Nearest Neighbors (KNN). Hyperparameter tuning and stratified cross-validation are used to optimize model performance.

3. Performance Evaluation

Standard predictive metrics—including accuracy, precision, recall, F1-score, and ROC-AUC—are computed. Subgroup-wise evaluations are also performed to highlight disparities across sensitive attributes like age, gender, and ethnicity.

4. Bias Detection and Fairness Metrics

Fairness-specific indicators are calculated, including:

- Statistical Parity Difference (SPD)
- Disparate Impact Ratio (DIR)
- Equalized Odds (EO)

  These metrics reveal whether the model's predictions are equitable across sensitive groups.

5. Visualization and Reporting

The tool generates confusion matrices, subgroup comparison charts, and prediction distribution plots, making results interpretable for both technical and non-technical users. Structured reports are then produced, summarizing dataset details, performance analysis, fairness evaluation, and recommended bias mitigation strategies (e.g., re-sampling, re-weighting, or post-processing adjustments).

This methodology ensures that predictive performance is not considered in isolation but alongside fairness, thus supporting ethical and transparent deployment of ML models in healthcare.

IV. RESULTS

Fig 1: ANACONDA PROMPT

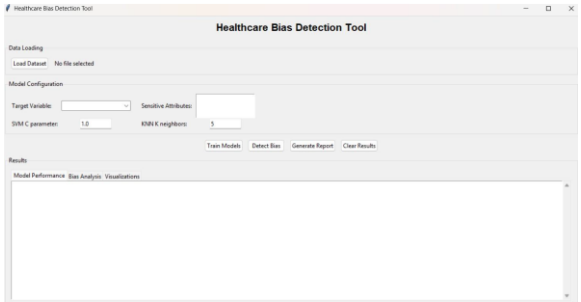

Fig 2: HOME PAGE

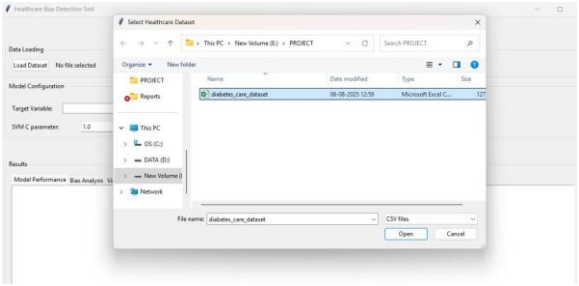

Fig 3: DATASET LOAD

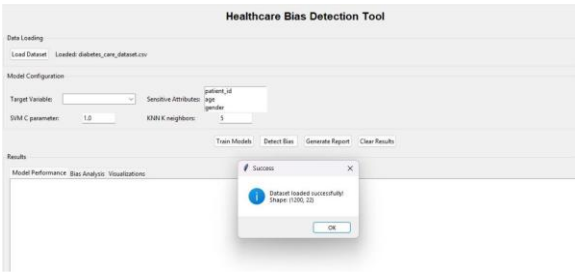

Fig 4: DATASET LOADED

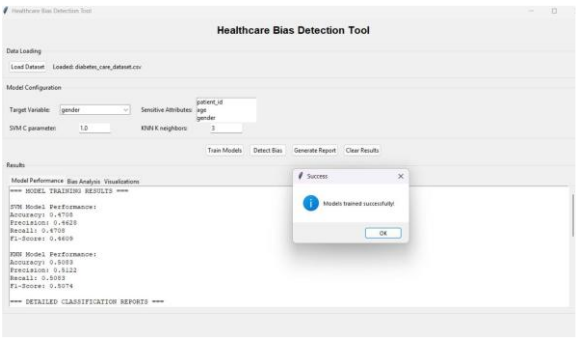

Fig 5: TRAIN MODELS

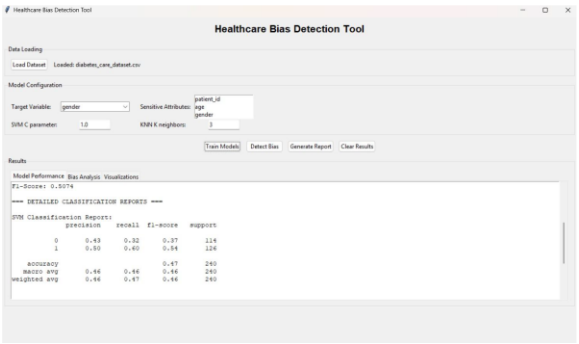

Fig 6: TRAIN MODELS OUTPUT

CONCLUSION

The clinical practice is to adopt an important step to detect significant steps in resolving moral and technical issues around the use of machine learning models. Unlike traditional assessment pipelines, which consider only overall accuracy, ensure that subgroup inequalities are neither deemed nor ignored by a combination of performance and

fairness metrics. The modular architecture of the system allows it to process model predictions, true labels and subgroups. It uses special engines to evaluate performance and fairness before producing outputs such as audit reports and visualisations such as output.

In addition to indicating the areas of prejudice, this method provides useful information that helps to make decisions about model signification or retreat. The system guarantees scalability, transparency and fertility using open-source tools and technologies such as the system python, pandas, skikit-learning and matplotlib. The most important thing is that it gives researchers, physicians and legislators the ability to identify and address prejudice, which promotes equity and trust in AI-managed healthcare. The suggested system provides a solid basis for developing fair, accountable and clinically reliable machine learning solutions, even though there are still issues with the real results connecting the fairness metrics and guaranteeing the availability of a representative dataset.

REFERENCES

[1] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.

[2] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dipping racial bias in an algorithm is used to manage population health," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[3] I. Y. Chen, F. D. Johansson, and D. Sontag, "Why is my classifier discriminatory?" In progress in nerve information processing systems *(NeurIPS)*, pp. 3543–3554, 2018.

[4] F. Kamiran and T. Calders, "Data preprocessing technology for classification without discrimination, "Knowledge and information system,, vol. 33, no. 1, pp. 1–33, 2012.

[5] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Proc. 20th Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, pp. 962– 970, 2017.

[6] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in Advances in Neural Information Processing Systems (NeurIPS*)*, pp. 3315–3323, 2016.

[7] F. Chen, C. Liu, and W. Lin, "Unmasking bias in artificial intelligence: A systematic review of EHR- based bias detection and mitigation," *Journal of Medical Systems*, vol. 48, no. 3, pp. 112–126, 2024.

[8] J. L. Cross, R. Mehta, and A. Singh, "Bias in medical artificial intelligence: Implications for clinical decision-making and fairness," *BMC Medical Ethics*, vol. 25, no. 1, pp. 1–15, 2024.

[9] Y. Huang, T. Zhang, and P. Li, "A Scoping view of fair machine learning techniques when applied to healthcare data," *Frontiers in Artificial Intelligence*, vol. 7, no. 122, pp. 1–18, 2024.

[10] R. Poulain, P. O'Connor, and J. Han, "Improving fairness in AI models on electronic health records: The case for federated learning methods," *arXiv preprint arXiv:2305.11386*, 2023.

[11] M. W. Sjoding, R. P. Dickson, T. J. Iwashyna, S. E. Gay, and T. S. Valley, "Racial bias in pulse oximetry measurement," *New England Journal of Medicine*, vol. 383, no. 25, pp. 2477–2478, 2020.

[12] J. Yang, A. Kumar, and F. Li, "An algorithmic bias training framework for mitigating adversarial in healthcare prediction models," *npj Digital Medicine*, vol. 6, no. 41, pp. 1–12, 2023.

[13] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Proc. 20th Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, vol. 54,pp. 962–970, 2017.

[14] Independent Review on Equity in Medical Devices, "Bias in pulse oximetry and clinical algorithms," UK Health Department Report, 2024.

[15] D. Ueda, Y. Chen, H. Kosaka, and T. Ichikawa, "Fairness of artificial intelligence in healthcare: A review and future directions," *Radiology: Artificial Intelligence*, vol. 5, no. 1, pp. 1–14, 2023.

[16] J. Cross, H. Yuan, and M. Schuchard, "Bias in medical AI: Implications for clinical decision-making," *NPJ Digital Medicine*, vol. 7, no. 66, pp. 1–11, 2024.

[17] S. Siddique, M. Khurram, and F. Jamil, "Survey on machine learning biases and

mitigation techniques," *Digital*, vol. 4, no. 1, pp. 101–120, 2024.

[18] A. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Machine learning of bias and fairness," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.

[19] J. Yang, R. Xie, A. Malhotra, and J. Park, "An Mitigating adversarial framework for algorithmic biases in clinical machine learning," *NPJ Digital Medicine*, vol. 6, no. 22, pp. 1–12, 2023.

[20] M. P. Cary Jr., C. Wang, and R. F. Brown, "Mitigating racial and ethnic bias and advancing equity in health care AI," *Health Affairs*, vol. 42, no. 3, pp. 329–338, 2023.

[21] D. Ueda *et al.*, "Fairness in intelligence in healthcare : review and future directions,"

[22] J. Yang *et al.*, "An Mitigating adversarial framework for algorithmic biases in clinical machine learning," *NPJ Digital Medicine*, 2023. Nature

[23] U.S. Drug and Food Administration, "Performance Evaluation of Pulse Oximeters (Executive summary/draft)," Feb 2024. Food and Drug Administration

[24] W. Ahmed, "Racial Biases Associated With Pulse Oximetry," *American Journal of Respiratory and Critical Care Medicine* (review), 2024. PubMed Central

[25] J. Carey, "Fairness in AI for healthcare," *ScienceDirect / review*, 2024. ScienceDirect

[26] Z. Obermeyer et al., "Dissecting racial bias of this algorithm used to manage the health of populations," *Science*, 2019. Science

[27] J. Yang et al., "An mitigating algorithmic biases in clinical machine learning of the algorithmic biases," *npj Digital Medicine*, 2023. Nature

[28] FDA, "Executive Summary: Performance Evaluation of Pulse Oximeters," Feb 2024 (draft/summary). Food and Drug Administration

[29] D. Ueda et al., "Fairness in healthcare: review and future directions," *Radiology: AI / PubMed*, 2024. PubMed

[30] J. Cross et al., "Bias in medical AI: Implications for clinical decision-making," *NPJ Digital Medicine*, 2024. PubMed Central

[31] N. Sourlos et al., "Recommendations for the creation of benchmark datasets for fairness research," 2024. PubMed Central