Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 20 No. 12 (2024) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v20i12.49997

PAPER

Review of Data Bias in Healthcare Applications

Atharva Prakash Parate[1], Aditya Ajay Iyer[2], Kanav Gupta[2], Harsh Porwal[2], P.C. Kishoreraja[1](✉), R. Sivakumar[3], Rahul Soangra[4]

[1]School of Computer Science and Information Systems, Vellore Institute of Technology, Vellore, Tamil Nadu, India

[2]School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India

[3]School of Electronics Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India

[4]Crean College of Health and Behaviour Sciences, Dale E and Sarah Ann Fowler School and Engineering, Chapman University, Irvine, CA, USA

kishoreraja.pc@vit.ac.in

ABSTRACT

In the area of medical artificial intelligence (AI), data bias is a major difficulty that affects several phases of data collection, processing, and model building. The many forms of data bias that are common in AI in healthcare are thoroughly examined in this review study, encompassing biases related to socioeconomic status, race, and ethnicity as well as biases in machine learning models and datasets. We examine how data bias affects the provision of healthcare, emphasizing how it might worsen health inequalities and jeopardize the accuracy of AI-driven clinical tools. We address methods for reducing data bias in AI and focus on different methods used for creating synthetic data. This paper explores several mitigating algorithms like SMOTE, AdaSyn, Fair-SMOTE, and BayesBoost. The optimized Bayesboost algorithm has been discussed. This approach showed more accuracy and addressed the error handling mechanism.

KEYWORDS

data bias, efficiency, fairness, artificial intelligence (AI), remedies, review

1 INTRODUCTION

In today's technology-driven world, data bias has emerged as a critical concern, especially in decision-making processes heavily reliant on data. Data bias refers to systematic errors or distortions in data that can lead to inaccurate or unfair outcomes. These biases can arise during data collection, analysis, interpretation, and publication [1], [2]. The implications of data bias are far-reaching, impacting various domains such as healthcare, criminal justice, finance, and social media. Attention to data bias has increased due to its potential to perpetuate discrimination, exacerbate societal inequalities, and compromise the fairness and integrity of data-driven systems. Anything that deviates from reality in data-related procedures is referred to as data bias. Errors or distortions in data gathering procedures, analytical strategies, frameworks for interpretation, and results publication fall under this category. These alterations may lead to incorrect inferences, which may have detrimental effects. Data bias, whether deliberate or inadvertent, can produce biased results and amplify preexisting prejudices and disparities in society [13]. Data bias persists and presents serious obstacles to academics, data scientists, legislators, and society

Parate, A.P., Iyer, A.A., Gupta, K., Porwal, H., Kishoreraja, P.C., Sivakumar, R., Soangra, R. (2024). Review of Data Bias in Healthcare Applications. *International Journal of Online and Biomedical Engineering (iJOE)*, 20(12), pp. 124–136. https://doi.org/10.3991/ijoe.v20i12.49997

Article submitted 2024-05-06. Revision uploaded 2024-07-19. Final acceptance 2024-07-19.

at large, despite efforts to eliminate it. When computers learn from big datasets to generate predictions or judgments, data bias has an especially noticeable effect on machine learning applications. The accuracy and representativeness of the training data are crucial factors that these algorithms rely on.

Data bias can arise at various stages of the data lifecycle [4], [10], [11], [13]–[18]. Selection bias occurs when certain groups are overrepresented or underrepresented in the dataset due to the sampling method used. For example, a survey conducted only online may not accurately represent the views of people without internet access. Response bias occurs when the data collected is influenced by the behavior or characteristics of the respondents, leading to skewed or inaccurate representations of reality. For instance, respondents may provide socially desirable answers rather than truthful ones, leading to biased results. Biases can also originate within algorithms themselves when they inadvertently learn and perpetuate discriminatory patterns present in the training data. Figure 1 shows different stages of bias. Algorithmic bias represents a significant issue that can lead to unjust or discriminatory outcomes. For example, a facial recognition algorithm trained on imbalanced datasets may struggle to accurately identify individuals from certain racial or gender groups [3], [50]. Likewise, lending algorithms might unintentionally exhibit favoritism toward specific demographic groups, leading to unequal access to financial services. To ensure fairness and equity in outcomes, addressing algorithmic bias [6], [9], [17], and [19] requires thorough scrutiny of both the training data and the design of the algorithm.

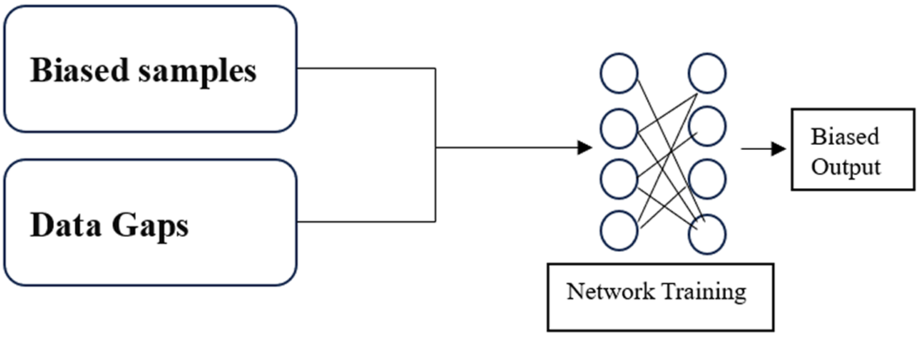

Fig. 1. Stages of data bias

Addressing data bias requires a multifaceted approach that encompasses both technical and ethical considerations [4], [17]. From a technical perspective, researchers and data scientists must implement robust data collection methods, carefully preprocess and clean the data, and employ bias mitigation techniques during algorithm training and evaluation. Data bias poses significant implications for justice, equity, and fairness in the era of big data and machine learning, presenting a substantial challenge. This research paper aims to review the synthesize the existing knowledge, identify the pitfalls of AI in dataset biasing, and suggest the direction for future work to mitigate the data bias [5], [7], and [11]. This paper underlines the critical importance of data bias in artificial intelligence (AI) systems.

2 METHOD

Artificial intelligence is rapidly transforming the healthcare landscape, offering exciting possibilities for disease diagnosis, treatment planning, and drug discovery. However, amidst this promise lies a hidden threat: bias. AI algorithms, like any human creation, are susceptible to inheriting and amplifying biases present in the data they are trained on. This can have serious consequences, particularly for marginalized communities, and exacerbate existing health inequities. One of the primary problems arises from biased datasets. If the data used to train AI models primarily reflects the health experiences of a specific demographic, the algorithm may become less accurate when encountering patients from different backgrounds. This can lead to misdiagnoses, particularly for conditions that present differently in various populations. Fear of bias in AI could also lead human healthcare professionals [5] to subconsciously downplay patient concerns, further hindering treatment. Mitigating bias in AI healthcare requires a multi-pronged approach [10], [11]. Firstly, ensuring diverse and representative datasets is crucial. This requires proactive efforts to collect data from all demographics and address historical underrepresentation in healthcare access. Secondly, ongoing monitoring and evaluation of AI algorithms for bias is vital. Human oversight remains essential, with healthcare professionals reviewing AI outputs and maintaining the final decision-making authority. Finally, promoting transparency and accountability in AI development is critical. There are two faces of bias of AI in healthcare [12]. One is data bias, and another is algorithmic bias [9]. However, additional biases may become apparent. These include AI clinician interaction and AI patient interaction.

2.1 Data bias in healthcare

Data bias arises from the inherent imperfections in the information used to train AI models [7]. These imperfections can stem from a variety of sources. If the training data primarily reflects the health experiences of a specific demographic (often the majority population), the AI may become less accurate when encountering patients from different backgrounds. This can lead to misdiagnoses, particularly for conditions that present differently in various populations. This comes from underrepresentation. Historical biases in healthcare data often reflect existing racial [3] and socioeconomic disparities [13]. Using such data can lead AI to perpetuate these biases. Imagine an algorithm trained on data where minority groups were less likely to receive certain screenings or treatments. This could lead the AI to downplay the seriousness of a condition in a minority patient based on skewed historical data. The data collection practices may lead to incomplete or inaccurate data, which distorts the reality AI models are trained on. Missing data from marginalized communities due to limited healthcare access further exacerbates the problem. This paper focuses on only seven types of data bias [7], [8], [10], and [12] in healthcare.

Scraped data bias. Scraped data bias in healthcare refers to the systematic errors and inaccuracies that arise when using data harvested from various sources like websites, social media platforms, or other online sources, leading to skewed or misleading outcomes in health-related studies and applications. This bias can stem from several factors, including the over-representation of certain demographics, outdated information, and the varying quality of data from different sources. When healthcare decisions are based on biased data, it can result in unequal treatment, misdiagnosis, and the perpetuation of existing health disparities [13]. Addressing scraped data bias requires meticulous data validation, the inclusion of diverse data sources, and the implementation of robust algorithms to ensure fair and accurate healthcare solutions.

Abstract data bias. Abstract data bias refers to biases introduced during the generation or extraction of data that may not be immediately evident. Scrutinizing data generation methods is essential to uncovering potential biases. For instance, if data is collected via surveys [20], biases may arise from respondent selection or question framing, leading to skewed representations of certain demographics or opinions. Addressing missing data or errors in abstract datasets is crucial for model performance and fairness. Techniques such as imputation or sensitivity analysis can help mitigate the impact of missing or erroneous data. Analyzing attribute distribution within abstract data can reveal biases that influence model outcomes, highlighting the importance of thorough data preprocessing. Example: In healthcare analytics, patient data extracted from electronic health records (EHRs) [21] might contain biases if certain demographic groups are more likely to seek medical care or have access to advanced healthcare facilities, influencing the model's ability to generalize across diverse populations.

Selection data bias. Selection bias occurs when the sample used in data collection is not representative of the entire population of interest, leading to skewed or non-generalizable insights. This bias can arise due to non-random sampling methods, inadequate sample sizes, or biased participant selection criteria. For example, if a study on smartphone usage only surveys [20] individuals who own high-end smartphones, the resulting data may not accurately reflect the broader population's smartphone usage patterns. This can lead to biased conclusions and unreliable predictive models. Identifying and mitigating selection bias [22] involves implementing rigorous sampling techniques and ensuring diverse representation within the dataset. Techniques such as stratified sampling or propensity score matching can help minimize the impact of selection bias on data analysis and modeling. Example: In healthcare studies, if clinical trials primarily enroll younger adults and exclude older populations, the effectiveness of treatments may be inaccurately assessed for older patients, leading to biased healthcare recommendations.

Survivorship data bias. Survivorship bias in healthcare occurs when conclusions are drawn from data that only includes patients who have survived a particular condition or treatment [23], ignoring those who have not. This is one type of selection bias. This bias can lead to overly optimistic assessments of treatment effectiveness and patient outcomes, as it disregards the experiences and outcomes of those who did not survive [24]. Consequently, healthcare policies and practices based on such biased data may fail to address the needs and challenges of all patients, particularly those at higher risk or with poorer prognoses. Mitigating survivorship bias requires a comprehensive analysis that includes both survivors and non-survivors to ensure a balanced and accurate understanding of healthcare outcomes.

Availability data bias. Availability bias refers to the tendency to rely heavily on information that is readily available or easily recalled when making judgments or decisions. In data analysis, availability bias can lead analysts to overemphasize recent or memorable data points, potentially overlooking less accessible but equally relevant information [25]. Addressing availability bias involves diversifying data sources and incorporating a broader range of information into analyses. Implementing structured decision-making processes that consider both recent and historical data can help mitigate the impact of availability bias on decision outcomes. For example, during public health crises [26], such as a pandemic, availability bias may lead policymakers to prioritize immediate responses based on recent outbreaks or media coverage, overlooking long-term health trends or alternative preventive measures.

Anchoring bias. Anchoring bias in healthcare occurs when clinicians rely too heavily on an initial piece of information—such as a patient's initial symptoms or a preliminary diagnosis—when making subsequent medical decisions. This cognitive bias [27] can lead to diagnostic errors, as the initial information unduly influences the interpretation of new evidence, potentially causing important symptoms to be overlooked or alternative diagnoses to be dismissed [28]. For example, if a doctor anchors on a diagnosis of a common condition, they might miss signs of a rarer but more serious illness. To combat anchoring bias, healthcare professionals need to remain open-minded, continually reassess initial assumptions, and consider a broad range of possibilities throughout the diagnostic and treatment process.

Interpretability bias. Interpretability bias in healthcare arises when complex medical data and models, such as those used in machine learning and AI, are interpreted incorrectly [29] due to a lack of transparency and understanding. This bias can result in misinformed decisions and actions because healthcare providers may rely on outputs they don't fully understand or that lack clear explanations. For instance, a predictive model might highlight a correlation without revealing the underlying causative factors, leading to potential misdiagnosis or inappropriate treatments. Addressing interpretability bias [30] requires developing and implementing models that provide clear, comprehensible insights, along with continuous education for healthcare professionals to enhance their understanding of these tools and their limitations.

Mitigating data bias in healthcare. Pre-processing, in-processing, and post-processing are the three basic categories into which bias reduction in machine learning models [2], [5], [7], [10], and [31] may be divided. Processing is a crucial step in addressing bias in machine learning models, as most bias is inherent in the data used for training. Pre-processing algorithms aim to reduce bias by manipulating the training data before training the algorithm, making it a conceptually simple yet effective approach. To reduce bias, several pre-processing methods are available, such as optimal data transformation, massaging, reweighing, and sampling. These strategies, which lessen bias and increase predictability in the model, can be as basic as basic data preparation approaches or as sophisticated as needed. Sampling and reweighing techniques are commonly used to adjust the balance of different groups present in the training data. For instance, if historical bias and discrimination are the root causes of bias in the data, relabeling or data transformation may be the best approach to decreasing the bias. On the other hand, if insufficient data, inconsistent data collection, or bad practices that are present in the data are the sources of bias, more complex pre-processing techniques such as optimized data transformation may be necessary. Hence, pre-processing is the most prevalent approach for bias mitigation. Key techniques in this category, such as resampling and reweighing, modify the training data distribution to address class or group imbalances. However, these methods have limitations in dealing with feature correlations and may result in data loss, making them less effective against confounding bias, algorithmic bias [9], and temporal bias. Other pre-processing strategies include data transformation to fill in missing data, relabeling, and domain adaptation, which provide early detection and mitigation of bias. Future research could explore the integration of multiple pre-processing methods in a single pipeline to effectively tackle various types of bias. In-processing approaches [32] for bias mitigation in machine learning models are a dynamic strategy that aims to reduce bias during model training. These approaches focus on adjusting the model's learning process to ensure fairness and prevent the model from simply learning to predict the majority class or the inherent bias in the training data.

Reweighing is one such method that modifies the weight of various training examples to concentrate more or less on underrepresented classes. By employing this tactic, the model is kept from picking up on the innate bias in the training set or the majority class. Transfer learning is another in-processing approach [34] that is especially beneficial for bias mitigation when data is limited. This method utilizes pre-trained models on large datasets to improve performance. By leveraging the knowledge gained from the pre-trained models, transfer learning can help reduce bias in the target model. Constraint optimization is another in-processing approach that enforces fairness constraints during learning. This method ensures that the model's predictions are fair and unbiased by imposing constraints on the learning process. Adversarial learning

[36] and [51] are a techniques used to assess the fairness of the training process. By training a model to learn fair representations of the data, adversarial learning can help reduce bias in the model's predictions. Hence, during model training, processing techniques offer dynamic bias reduction measures. Reweighting is one way to stop the model from simply learning to predict the majority class or the inherent bias in the training data. It does this by adjusting the relevance of various training samples to focus more on underrepresented classes or less on overrepresented ones. Transfer learning uses vast datasets of pre-trained models to improve performance, making it particularly useful for mitigating bias when data is scarce. Additional processing methods for machine learning models include adversarial learning, which evaluates the training's fairness, regularization, which prevents overfitting, and constraint optimization, which imposes fairness restrictions during learning. Post-processing methods, which modify model output, offer a viable method for reducing bias in machine learning algorithms [33]. These techniques provide an adaptive and versatile way to assess fairness, especially in black-box AI algorithms. Post-processing approaches can assist in correcting any bias produced during the training phase by altering the output. Calibration is one such post-processing method that fixes bias in probabilities that are expected. This technique modifies a model's projected probability to make sure it is reasonable and correct. Calibration is particularly useful for models that are used to make decisions based on predicted probabilities, such as in credit scoring or medical diagnosis [35], [36]. Another post-processing technique is decision threshold selection. This method involves adjusting the threshold for classifying examples as positive or negative to ensure that the model's predictions are fair and unbiased. Decision threshold selection can be particularly useful for models that are used to make decisions that have significant consequences, such as in criminal justice or hiring decisions. As a result, just one study in this evaluation used postprocessing techniques, which modify model output. These techniques are currently underused. Several additional possible postprocessing techniques, including decision threshold selection and calibration to address bias in projected probabilities, have been brought to light in previous surveys as ways to reduce bias in machine learning models. Subsequent investigations ought to focus on creating resilient postprocessing techniques that may efficiently identify, reduce, and elucidate prejudice, substantially enhancing equity in AI-related electronic health record (EHR) systems.

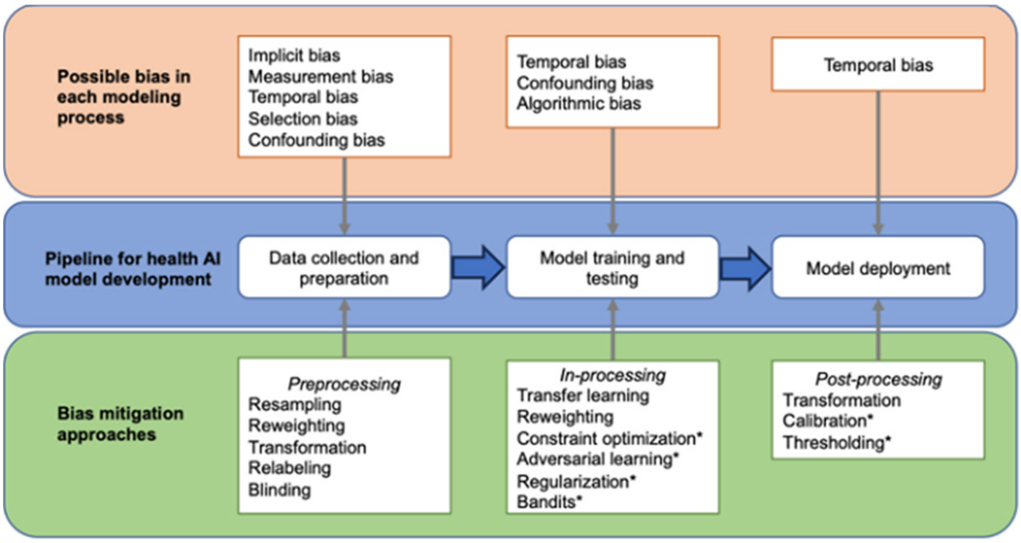

Fig. 2. Bias mitigation approaches

There are several reasons for the performance gap observed in AI models. Bias can be introduced at various stages of model construction, including data collection, preparation, model development, assessment, and deployment in clinical settings. For example, the algorithm might have been trained primarily on data from white patients, or access to medical information for black individuals might be more challenging. Additionally, the method used to train the model to predict risk is likely influenced by underlying societal disparities [13] in healthcare access and spending. Regardless of the cause, an algorithm that disproportionately assigns false negatives could lead to fewer follow-up scans, potentially resulting in more undiagnosed and untreated cancer cases and exacerbating health disparities for already disadvantaged groups. Figure 2 shows the bias-mitigating approaches in healthcare AI.

Addressing the data bias with synthetic data. Even with the greatest of intentions, obtaining unbiased datasets can be difficult. Underrepresented groups may find it difficult to participate in data-sharing programs due to worries about privacy, anonymity, and trust. Therefore, it is crucial to employ methods for identifying and mitigating bias to improve healthcare outcomes. One of the approaches to addressing data bias is generating synthetic datasets [37], which accurately represent diverse patient populations. BayesBoost is a powerful technique for identifying underrepresented groups and addressing biases within datasets. We explore the application of BayesBoost in generating synthetic data [38], [39], focusing on its use within the UK's Clinical Practice Research Datalink (CPRD). Synthetic data offers numerous advantages over real patient data for statistical analysis, machine learning, and AI research, particularly in identifying and mitigating biases in ground truth datasets [40]. While conventional methods such as SMOTE and AdaSyn [41], [42] have been effective in balancing classes in datasets with imbalances, they can compromise fairness by not considering sensitive characteristics during the equalization process. Fair-SMOTE addresses these limitations by balancing data based on both sensitive and class features, although it may not fully capture the true data distribution. In contrast, BayesBoost [44] has shown promise in creating synthetic data that closely mirrors the original data distribution. This has led to improved accuracy and performance metrics in various experiments, highlighting its potential as a valuable tool in synthetic data generation for research and analysis. Among the various techniques used for synthetic data generation, BayesBoost has garnered attention for its effectiveness in identifying underrepresented groups and mitigating biases within datasets. Developed as an extension of traditional boosting algorithms, BayesBoost leverages Bayesian inference to generate synthetic data that closely resembles the distribution of real data. By incorporating information about class labels, sensitive attributes, and target diseases, BayesBoost can produce synthetic datasets that accurately reflect the diversity and complexity of patient populations. The goal of investigating BayesBoost in the context of medical research is to enhance the decision-support systems and prediction models [40], [44] generalizability, fairness, and accuracy. Conventional methods for handling unbalanced datasets, including (AdaSyn) adaptive synthetic sampling and (SMOTE) synthetic minority over-sampling technique), might not fully account for the influence of sensitive variables on model performance or accurately represent the underlying data distribution. Although Fair-SMOTE balances data according to sensitive qualities in addition to class labels to overcome these restrictions, it might not be able to effectively represent real data distribution. Identifying underrepresented groups in databases is a critical initial step, especially in the creation of synthetic data. BayesBoost has shown promise in identifying and mitigating biases in data [45], thereby improving learning outcomes. This approach has proven highly beneficial for synthetic dataset services, such as those offered by the Clinical Practice Research Datalink in the UK [46].

For advanced statistical analysis, machine learning, and AI research applications, the use of synthetic data offers several advantages over real patient data. One of the primary benefits is the ability to identify and reduce biases in ground truth datasets. Unlike biased ground-truth datasets, synthetic data remains unaffected by inaccurate correlations, distributions, or structurally missing data. While traditional strategies such as AdaSyn and SMOTE improve model performance by balancing classes, they may inadvertently undermine equity by randomly selecting and adjusting the qualities of two groups. Fair-SMOTE, on the other hand, balances data based on class and sensitive properties, addressing the shortcomings of SMOTE and AdaSyn. This ensures that the data contains an equal number of positive and negative instances for both affluent and underprivileged groups. However, this approach may yield data findings that do not accurately reflect the actual data distribution, despite being highly effective in reducing bias and enhancing fairness. In contrast, BayesBoost has been shown to generate data that closely resembles the original data distribution. This was observed in simulation experiment results comparing ground truth data with BayesBoost-generated data. Results from using SMOTE, AdaSyn, and Fair-SMOTE showed similar performance values for COVID-19 and cardiovascular disease (CVD) data. However, datasets produced using BayesBoost [47] demonstrated higher accuracy values compared to those produced using SMOTE, AdaSyn, and Fair-SMOTE. This is an improved version of BayesBoost algorithm using stratified sampling, which first divides the dataset Bias into separate training and validation subsets. A new dataset is formed, which highlights potential areas of bias or data imbalance that require further investigation and mitigation. Error handling mechanisms are incorporated to address challenges encountered during synthetic data generation [48], [49], ensuring the robustness and integrity of the enhanced dataset. Throughout the algorithm's implementation, clarity and readability are emphasized through logical organization and concise documentation. By prioritizing clarity, the algorithm becomes accessible to practitioners seeking to address data biases in machine learning applications [52], fostering transparency and understanding throughout the bias mitigation process. The BayesBoost algorithm mitigates the data bias in healthcare. The optimized BayesBoost algorithm represents a comprehensive approach to bias mitigation, integrating stratified sampling, modular training, uncertainty analysis, feature prioritization, synthetic data generation, error handling, and clarity in implementation. This holistic framework enables practitioners to develop fairer and more accurate machine learning models, ultimately promoting fairness and transparency in data-driven decision-making. So, the detection of underrepresented groups of patients and the generation of synthetic data are critical aspects of bias mitigation in machine learning models. Techniques such as BayesBoost offer promising solutions to these challenges, leading to improved learning outcomes and more equitable AI applications. Future research should continue to explore and refine these techniques to further enhance their effectiveness and applicability.

3 RESULTS AND DISCUSSION

The extensive review of current research on the data bias of AI in healthcare exposes significant advancements in employing different mitigating methods for data bias, which helps AI-centric clinical and patient healthcare systems. Key findings from the literature indicate:

- Impact of different types of data bias in AI in healthcare: Different types of data bias in healthcare are discussed. It has been found that selection bias is very severe and affects the entire AI system, leading to poor generalization across diverse patient populations. This can result in systemic inequalities in healthcare delivery. It may be underrepresented in the training data.
- Mitigating the data bias using synthetic datasets: Several mitigating algorithms like SMOTE, AdaSyn, Fair-SMOTE, and BayesBoost generate synthetic datasets. It has been observed that Bayboost showed the best result among others with accuracy values compared to those produced using SMOTE, AdaSyn, and Fair-SMOTE.
- Using the BayesBoost algorithm: BayesBoost excels at finding hidden groups in data, especially when dealing with sensitive information and a specific disease. It tackles data bias by generating realistic synthetic data that closely mirrors real-world data, reducing inconsistencies. The optimized version of the BayesBoost algorithm has been discussed using the stratified sampling method. This approach gives more accuracy and also addresses the error handling mechanisms.

4 CONCLUSION

While AI holds immense potential for healthcare advancements, data bias within these applications can exacerbate existing healthcare disparities. To ensure fair and effective use of AI, we must address bias throughout the development process, from ensuring diverse training data to implementing robust evaluation methods. This research paper discusses the different types of data bias in AI in healthcare applications and finding the severity of each data bias. Different mitigation methods for data bias have been deliberated. We explored the application of BayesBoost in generating synthetic data and compared it with SMOTE, AdaSyn, and Fair-SMOTE. We found that BayesBoost showed high accuracy for underrepresented groups of patients and the generation of synthetic data. The improved version of BayesBoost algorithm showed greater accuracy and addressed the error handling mechanisms. Techniques like BayesBoost offer promising solutions to these challenges, leading to improved learning outcomes and more equitable AI applications. Future research should continue to explore and refine these techniques to further enhance their effectiveness and applicability. Only through proactive mitigation strategies can we harness the true potential of AI to improve health outcomes for all.

5 REFERENCES

[1] S. Siddique, M. A. Haque, R. George, K. D. Gupta, D. Gupta, and M. J. H. Faruk, "Survey on machine learning biases and mitigation techniques," *Digital*, vol. 4, no. 1, pp. 1–68, 2024. https://doi.org/10.3390/digital4010001

[2] E. Ferrara, "Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies," *Science*, vol. 6, p. 3, 2024. https://doi.org/10.3390/sci6010003

[3] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019. https://doi.org/10.1126/science.aax2342

[4] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, and A. Tzovara, "Addressing bias in big data and AI for health care: A call for open science," *Perspective*, vol. 2, no. 10, pp. 1–9, 2021. https://doi.org/10.1016/j.patter.2021.100347

[5] D. Ueda *et al.*, "Fairness of artificial intelligence in healthcare: Review and recommendations," *Japanese Journal of Radiology*, vol. 42, pp. 3–15, 2023. https://doi.org/10.1007/s11604-023-01474-3

[6] M. H. Chin *et al.*, "Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care," *JAMA Network Open*, vol. 6, no. 12, p. e2345050, 2023. https://doi.org/10.1001/jamanetworkopen.2023.45050

[7] M. Mittermaier, M. M. Raza, and J. C. Kvedar, "Bias in AI-based models for medical applications: Challenges and mitigation strategies," *NPJ Digital Medicine Nature*, vol. 6, 2023. https://doi.org/10.1038/s41746-023-00858-z

[8] N. Shahbazi, Y. Lin, A. Asudeh, and H. V. Jagadish, "Representation bias in data: A survey on identification and resolution techniques," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–39, 2023. https://doi.org/10.1145/3588433

[9] R. K. E. Bellamy *et al.*, "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, nos. 4/5, pp. 4:1–4:15, 2019. https://doi.org/10.1147/JRD.2019.2942287

[10] K. N. Vokinger, S. Feuerriegel, and A. S. Kesselheim, "Mitigating bias in machine learning for medicine," *Communications Medicine*, vol. 1, no. 1, 2021. https://doi.org/10.1038/s43856-021-00028-w

[11] Agnieszka *et al.*, "A survey on bias in machine learning research," *Knowledge-Based Systems*, 2023.

[12] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, and A. Tzovara, "Addressing bias in big data and AI for health care: A call for open science," *Patterns Journal*, vol. 2, no. 10, p. 100347, 2021. https://doi.org/10.1016/j.patter.2021.100347

[13] L. A. Celi *et al.,* "Sources of bias in artificial intelligence that perpetuate healthcare disparities," *PIOS Digital Health*, vol. 1, no. 3, pp. 1–19, 2022. https://doi.org/10.1371/journal.pdig.0000022

[14] S. Siddique, M. A. Haque, R. George, K. D. Gupta, D. Gupta, and M. J. H. Faruk, "Survey on machine learning biases and mitigation techniques," *Digital*, vol. 4, no. 1, pp. 1–68, 2024. https://doi.org/10.3390/digital4010001

[15] QuestionPro, "Data bias: Identifying and reducing in surveys and analytics." https://www.questionpro.com/blog/data-bias/

[16] P. Krishnamurthy *et al.*, "Understanding data bias: Types and sources of data bias," *Towards Data Science*, 2019. https://towardsdatascience.com/survey-d4f168791e57

[17] S. Silva and M. Kenney, "Algorithms, platforms, and ethnic bias: An integrative essay," Berkeley Roundtable on the International Economy, University of California, Berkeley, Reports, 2018.

[18] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2019. https://doi.org/10.1145/3457607

[19] S. Akter, G. McCarthy, S. Sajib, K. Michael, Y. K. Dwivedi, J. D'Ambra, and K. N. Shen, "Algorithmic bias in data-driven innovation in the age of AI," *International Journal of Information Management*, vol. 60, p. 102387, 2021. https://doi.org/10.1016/j.ijinfomgt.2021.102387

[20] Ali Akbar Jamali, Corinne Berger, and Raymond J. Spiteri, "Identification of depression predictors from standard health surveys using machine learning," *Current Research in Behavioral Sciences,* vol. 7, p. 100157, 2024. https://doi.org/10.1016/j.crbeha.2024.100157

[21] B. Al-Sahab, A. Leviton, T. Loddenkemper, N. Paneth, and B. Zhang, "Biases in electronic health records data for generating real-world evidence: An overview," *Journal of Healthcare Informatics Research*, vol. 8, pp. 121–139, 2024. https://doi.org/10.1007/s41666-023-00153-2

[22] V. K. Chauhan *et al.*, "Sample selection bias in machine learning for healthcare," Preprint, Institute of Biomedical Engineering, University of Oxford, UK, 2024.

[23] E. Mark Czeisler *et al.*, "Uncovering survivorship bias in longitudinal mental health surveys during the COVID-19 pandemic," *Epidemiology and Psychiatric Sciences*, vol. 30, p. e45, 2021. https://doi.org/10.1017/S204579602100038X

[24] "Healthcare research: Analyzing survivorship bias risk in clinical studies," Faster Capital, 2024.

[25] P. Li, Z. Y. Cheng, and G. L. Liu, "Availability bias causes misdiagnoses by physicians: Direct evidence from a randomized controlled trial," *Internal Medicine*, vol. 59, no. 24, pp. 3141–3146, 2020. https://doi.org/10.2169/internalmedicine.4664-20

[26] Kwaku Kyere, Taiwo O. Aremu, and Oluwafemi A. Ajibola, "Availability bias and the COVID-19 pandemic: A case study of legionella pneumonia," *Cureus*, vol. 14, no. 6, p. e25846, 2022. https://doi.org/10.7759/cureus.25846

[27] G. Saposnik, D. Redelmeier, C. C. Ruff, and P. N. Tobler, "Cognitive biases are associated with medical decisions," *BMC Medical Informatics and Decision Making Journal*, vol. 16, no. 1, 2016. https://doi.org/10.1186/s12911-016-0377-1

[28] M. H. Elizabeth Hammond *et al.*, "Bias in medicine: Lessons learned and mitigation strategies," *JACC: Basic to Translational Science*, vol. 6, no. 1, pp. 78–85, 2021. https://doi.org/10.1016/j.jacbts.2020.07.012

[29] C. Meng, L. Trinh, N. Xu, J. Enouen, and Y. Liu, "Interpretability and fairness evaluation of deep learning models on the MIMIC-IV dataset," *Scientific Reports*, vol. 12, no. 1, 2022. https://doi.org/10.1038/s41598-022-11012-2

[30] H. Hakkoum, I. Abnane, and A. Idri, "Interpretability in the medical field: A systematic mapping and review study," *Applied Soft Computing*, vol. 117, p. 108391, 2022. https://doi.org/10.1016/j.asoc.2021.108391

[31] A. Balayn, C. Lofi, and G. J. Houben, "Managing bias and unfairness in data for decision support: A survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems," *The VLDB Journal*, vol. 30, pp. 739–768, 2021. https://doi.org/10.1007/s00778-021-00671-8

[32] M. Wan, D. Zha, N. Liu, and N. Zou, "In-Processing modeling techniques for machine learning fairness: A survey," *ACM Transactions on Knowledge Discovery from Data*, vol. 17, no. 3, pp. 1–27, 2023. https://doi.org/10.1145/3551390

[33] K. Pranay Lohia *et al.*, "Bias mitigation post-processing for individual and group fairness," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2019)*, Brighton, UK, 2019, pp. 2847–2851. https://doi.org/10.1109/ICASSP.2019.8682620

[34] Z. Chen, J. M. Zhang, F. Sarro, and M. Harman, "A comprehensive empirical study of bias mitigation methods for machine learning classifiers," *ACM Transactions on Software Engineering and Methodology*, vol. 32, no. 4, pp. 1–30, 2023. https://doi.org/10.1145/3583561

[35] R. Roelofs, N. Cain, J. Shlens, and M. C. Mozer, "Mitigating bias in calibration error estimation," in *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Valencia, Spain, 2022, vol. 151, pp. 1–19.

[36] M. Hort, Z. Chen, J. M. Zhang, M. Harman, and F. Sarro, "Bias mitigation for machine learning classifiers: A comprehensive survey," *ACM Journal on Responsible Computing*, vol. 1, no. 2, pp. 1–52, 2024. https://doi.org/10.1145/3631326

[37] N. Patki, R. Wedge, and K. Veeramachaneni, "The synthetic data vault," in *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2016, pp. 399–410. https://doi.org/10.1109/DSAA.2016.49

[38] Z. Wang, P. Myles, and A. Tucker, "Generating and evaluating synthetic UK primary care data: Preserving data utility and patient privacy," in *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, UK, 2019, pp. 126–13. https://doi.org/10.1109/CBMS.2019.00036

[39] A. Tucker, Z. Wang, Y. Rotalinti, and P. Myles, "Generating high-fidelity synthetic patient data for assessing machine learning healthcare software," *NPJ Digital Medicine*, vol. 3, 2020. https://doi.org/10.1038/s41746-020-00353-9

[40] B. Draghi, Z. Wang, P. Myles, and A. Tucker, "Identifying and handling data bias within primary healthcare data using synthetic data generators," *Heliyon*, vol. 10, no. 2, 2024. https://doi.org/10.1016/j.heliyon.2024.e24164

[41] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2009, pp. 1322–1328.

[42] V. Nitesh Chawla, W. Kevin Bowyer, O. Lawrence Hall, and W. Philip Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. https://doi.org/10.1613/jair.953

[43] B. Draghi, Z. Wang, P. Myles, and A. Tucker, "BayesBoost: Identifying and handling bias using synthetic data generators," *SSRN*, 2022. https://doi.org/10.2139/ssrn.4052302

[44] R. González-Sendino, E. Serrano, and J. Bajo, "Mitigating bias in artificial intelligence: Fair data generation via causal models for transparent and explainable decision-making," *Future Generation Computer Systems*, vol. 155, pp. 384–401, 2024. https://doi.org/10.1016/j.future.2024.02.023

[45] T. Yang, C. Han, C. Luo, P. Gupta, J. M. Phillips, and Q. Ai, "Mitigating exploitation bias in learning to rank with an uncertainty-aware empirical bayes approach," in *Proceedings of the ACM on Web Conference 2024 (WWW '24)*, 2024, pp. 1486–1496. https://doi.org/10.1145/3589334.3645487

[46] Clinical Practice Research Datalink, "CPRD COVID-19 symptoms and risk factors synthetic dataset April 2021," 2021. https://doi.org/10.48329/fbjh-es87

[47] Z. Wang, B. Draghi, Y. Rotalinti, D. Lunn, and P. Myles, "High-fidelity synthetic data applicaztions for data augmentation," *Deep Learning: Recent Findings and Research*, 2024. https://doi.org/10.5772/intechopen.113884

[48] S. Hao, W. Han, T. Jiang, Y. Li, and H. Wu, "Synthetic data in AI: Challenges, applications, and ethical implications," Huazhong University of Science and Technology, China, *arXiv Pre print arXiv:2401.01629*, 2024.

[49] Q. H. Nguyen, T. T. Vu, A. T. Tran, and K. Nguyen, "Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation," in *Thirty-Seventh Conference on Neural Information Processing Systems,* China, 2023.

[50] D. Cirillo *et al.*, "Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare," *NPJ Digital Medicine*, vol. 3, no. 1, 2020. https://doi.org/10.1038/s41746-020-0288-5

[51] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES '18*, USA, 2018, pp. 335–340. https://doi.org/10.1145/3278721.3278779

[52] J. Chakraborty, S. Majumder, and T. Menzies, "Bias in machine learning software: Why? How? What to do?" in *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*, 2021, pp. 429–440. https://doi.org/10.1145/3468264.3468537

6 AUTHORS

Atharva Prakash Parate, Aditya Ajay Iyer, Kanav Gupta, and **Harsh Porwal** are students at the Department of Information Technology, School of Computer Science and Information Systems, and School of Computer Science and Engineering, VIT University, Vellore (E-mail: gatharvaprakash.parate2021@vitstudent.ac.in, adityaajay.iyer2021@vitstudent.ac.in, kanav.gupta2021@vitstudent.ac.in, harsh.porwal2021@vitstudent.ac.in).

P.C. Kishoreraja is a Professor at the School of Computer Science and Information Systems at Vellore Institute of Technology (VIT), Vellore, India. His research interests omclude Machine Learning Algorithms and Internet of Things (E-mail: kishoreraja.pc@vit.ac.in).

R. Sivakumar is a Professor at the Division of Sensors and Biomedical Technology, School of Electronics Engineering, VIT University, Vellore. He researches in Signal Processing, Image Processing and Biomedical Engineering. (E-mail: rsivakumar@vit.ac.in).

Rahul Soangra is an Assistant Professor of physical therapy at the Crean College of Health and Behavioral Sciences at Chapman University. He is also an adjunct faculty in the Fowler School of Engineering. His research interests are fall risk assessment using wearable inertial sensors, machine learning based classification of gait in idiopathic toe walkers, and fall intervention in older adults and stroke patients.