International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Bias Detection and Mitigation within Decision Support System: A Comprehensive Survey

Jyoti Prakhar[1] and Md. Tanwir Uddin Haider (SMIEEE)[2]

Submitted: 25/04/2023 **Revised**: 27/06/2023 **Accepted**: 07/07/2023

Abstract: In decision support system biases plays a vital role to lead unfair and discriminatory outcomes within the system, which can have serious consequences for people and society as a whole. While significant work has been done in classical machine learning and deep learning to address these difficulties, still there is a need for extensive surveys that evaluate various real-world applications and causes of bias in decision support systems. In this paper, we present a comprehensive survey that explores the biases, detection of biases, mitigation of biases, and fairness metrics to measure the degree of fairness in decision support systems. Further, we also identified several challenges related to biases such as minimizing biases when working with inadequate datasets, ensuring proper representation of protected attributes, developing efficient and direct methods for bias detection, identifying effective approaches for mitigating biases at various stages of the model, developing strategies to effectively mitigate multiple biases in the system to build a fair prediction model and at last, exploring and refining fairness metrics to achieve more fair results. We have also provided the research questions based on these challenges with the solutions and interesting future research avenues that might help to alleviate the problem of bias in decision support systems. We hope that this poll will motivate scholars to confront these issues and help the creation of more equitable systems.

Keywords: *Bias, Decision support system, Fairness Metrics, Mitigation.*

1. Introduction

Decision Support Systems (DSS) are rapidly being utilized in a variety of sectors to help people make quicker and more intelligent decisions [1]. However, these systems are increasingly employed in various industries and domains, and concerns about biases and unfairness in their outputs have emerged. While machine learning techniques have sought to address these concerns, bias in training data remains a substantial underlying cause of unfairness, and standard algorithmic approaches are incapable of completely mitigating these flaws. There is a pressing need for reliable and resilient approaches that can overcome data biases to improve the accuracy of DSS models.

The decision-support system focuses on accessing and manipulating a temporal series of organizational internal as well as external data. For example, such systems can assist in deciding which applications to select for a job (like Amazon [2] does) and in determining how to react in the event of an accident in a self-driving car [3]. DSS with Online Analytical Processing (OLAP) provides the highest level of capacity and decision support for extensive historical data analysis [4]. An effective DSS emphasizes the creation of data representations and helps to ensure that proper data is obtained. The decision support systems have gotten increasingly complex as technology has progressed. It has some important features such as ad-hoc data filtering and retrieval, alerts and triggers, data summarization, metadata creation and retrieval, and statistical analysis [5]. Nowadays, DSS is applied in various applications such as loan prediction, job search, medical conditions, etc. Besides the many advantages of these systems in different scenarios, a DSS has many ethical concerns, such as: (i) the DSS is frequently accused of lacking clarity; (ii) their results are frequently unaccountable; (iii) they may violate the confidentiality of many stakeholders; and (iv) they are frequently accused of being unfair to certain groups of the population, i.e., biases in the system.

Biases in decision support systems can have serious effects, such as discriminating outcomes, distorted recommendations, and promoted social disparities. As a result, recognizing and correcting biases in these systems has emerged as an essential goal for researchers, practitioners, and policymakers together. This survey gives an overview of biases, detection of biases, mitigation of biases, and highlights the various fairness metrics that are used to measure the degree of fairness. Based on the survey, we have also provided the major challenges such as minimizing biases when working with inadequate datasets, ensuring proper representation of protected attributes, developing efficient and direct methods for bias detection, establishing a standardized approach for bias mitigation, developing strategies to effectively mitigate multiple biases in the system to build a fair prediction model and at last, exploring and refining fairness metrics to achieve more fair results. We have also framed the research questions based

1Dept. of CSE, National Institute of Technology Patna, Bihar – 800005, India
Email: jyotip.ph21.cs@nitp.ac.in
2 Dept. of CSE, National Institute of Technology Patna, Bihar – 800005, India
Email: tanwir@nitp.ac.in
** Corresponding Author Email: jyotip.ph21.cs@nitp.ac.in*

on the identified gaps with the solutions. Our goal is to give a comprehensive overview of the present methods, methodologies, and strategies for detecting and mitigating biases in these systems. We intend to shed light on the problems and opportunities related to bias identification and reduction by analyzing the present state of research and practice.

Problem Focus

In this paper, we focus on exploring different scenarios for biases with detection, mitigation, and fairness metrics of the biases within the DSS.

Motivation

Data mining and machine learning researchers have begun to address unfairness or bias from various perspectives, such as by assessing trained model outputs, reducing unfairness by post-processing the system's outputs, or pre-processing the training data [6]. However, many of these initiatives fail to address the underlying causes of unfair systems. Moreover, it is stated that they are difficult to obtain and apply by professionals in real-life situations. That is why we feel that the data management community should do more detailed work on the biases, and further research in that direction.

Contribution

The purpose of this survey is to focus on the interest of the researcher by outlining the different biases that can exist in the system and the data management community educates the public about biases in DSS. We have also discussed sources of biases and fairness metrics that can be applied to data management as a part of a decision support system, if addressed, should move systems in the direction of a fair state. We also focused on different research questions and proposed solutions based on the relevant challenges.

To the best of our knowledge, this is the first survey to systematically and comprehensively cover bias, detection of bias, mitigation of bias, and fairness metrics. In summary, our contributions include the following:

- We figure out the different biases that can occur in a DSS with the sources of bias and also described the techniques for the detection of biases, mitigation of biases, and fairness metrics that can be applied to the system (Sect. 2).
- We systematically surveyed various research papers related to biases, detection of biases, mitigation of biases, and fairness metrics that can occur in a decision support system. Based on the survey performed in (Sect. 3), we identified several challenges related to biases, detection of biases, mitigation of biases, and fairness metrics where research can be carried out.
- Finally, we identify the research questions based on these challenges and propose solutions to these questions, along with related future work. (Sects. 4 and 5).

The remaining paper is organized as follows:

In section 2, the background information includes taxonomy and sources of biases with various methods for the detection of biases, mitigation of biases, and different fairness metrics. There is a thorough literature review in Section 3. The research publications that deal with bias are included in section 3.1. Research studies related to the detection 3 of bias are covered in Section 3.2. Research papers related to the mitigation of biases are included in section 3.3. And at last, the assessment of research publications that examine fairness measures in section 3.4 and derived the challenges with research questions from the identified research gaps. The findings with the solution to the research questions stated in this work are presented in Section 4. Finally, Section 5 concludes the paper by summarizing the key findings and offering suggestions for future research directions in this field.

2. Background

In this section, we provide background knowledge on biases, detection of biases, mitigation of biases, and different fairness metrics that can be applied to the system.

2.1. Understanding of Biases

Bias is a logical concept or assumption that limits one's ability to make reasonable judgments based on knowledge and inquiry i.e., Bias in any system can lead the system to behave in a specific manner, which degrades the working ability of the respective system. Biases can be broadly defined as statistical and social. Several types of bias, including sentimental bias and observational bias, might appear in the system under different circumstances. These biases can also occur in any phase of the decision support system, which is a part of artificial intelligence namely, Data Creation, Problem Formulation, Data Analysis, Validation, and Testing. In these phases, different types of biases can be present. Figure 1 gives a visual representation of the bias taxonomy as it applies to the various phases of the AI process [7]:

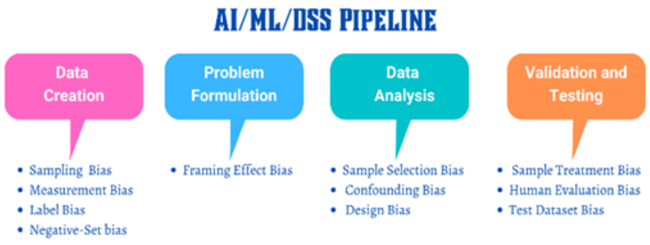


Fig. 1. Classification of bias

Biases can occur in the data Creation phase such as sampling bias, measurement bias, label bias, and negative-set bias. The sampling bias results when a certain collection of instances is frequently used to construct datasets [8] [9]. The measurement bias results from human measurement errors or other inherently human behaviors during data collecting. The label bias that occurs due to inconsistencies in the labeling process is referred to as label bias [10] [11] [12]. And the negative-set bias is introduced into a dataset because of a lack of samples indicative of the rest of the world. Bias can also arise during the problem formulation phase can arise such as framing effect bias. The framing effect bias occurs when the information presented affects a person's choice among several options more than the information itself [13]. Similarly in the data analysis phase, there are numerous ways in which biases may emerge in the algorithm or during data analysis such as sample selection bias, confounding bias, and design-related bias. The sample selection bias happens when persons, groups, or data are selected for research in such a manner that the samples are not representative of the population under investigation [14]. Confounding bias may develop if the algorithm learns the improper relationship by using partial data or if it ignores the suitable relationships between the intended output and characteristics [15] [16]. The design-related bias, also known as algorithm bias, may occur owing to the limitations of algorithmic outputs or other system limits such as computer capabilities [17] [18] [19]. The last stage of DSS is the validation and testing phase in which different types of biases can occur which are sample treatment bias, human evaluation bias, and test dataset bias. The sample treatment bias is established during the process of selecting and treating a particular group of persons for a form of therapy [20]. Human evaluation bias arises due to human evaluators required to verify an AI model's performance; these also contribute to biases as well [12]. The test dataset bias occurs when the validation and test datasets may potentially contain biases due to sample selection and label errors [8].

2.2 Sources of Biases

Biases in a DSS can be caused by a variety of factors such as data adequacy, data bias, and model adequacy. In Data adequacy, the system becomes biased when the data are insufficient to adequately represent diverse groups. Furthermore, in data bias, when the data that is currently available does not accurately depict the population, biases may arise. Lastly, the choice of qualities to include in the model may introduce bias. And in the model adequacy, bias may occur because the model design favors certain unique groupings over others. For instance, a linear model may favorably characterize one group over another.

These biases can lead to the degradation of the overall performance of the system so there is a need for the detection of biases to improve the system's decisions making capabilities. The various techniques are employed for the detection of biases which is discussed in section 2.3.

2.3 Detection of Biases

The detection of biases refers to the process of finding and revealing biases that may exist in a variety of situations, including data, algorithms, decision-making processes, and many others. It entails the inspection, analysis, and critical assessment of data, information, or behaviors to discover any bias, prejudice, distortion, or injustice that may exist. The purpose of bias detection is to raise awareness of these biases, understand their effect, and strive towards minimizing their influence to assure justice, objectivity, and equality in decision-making and results. Recently, various techniques have been widely used for the detection of biases such as the BERT model, self-supervised learning, unsupervised learning, and many more. Here, we will discuss some of the major techniques for the detection of biases in detail.

The BERT model includes comparing the word frequency distributions across various corpora to find biased language patterns. These corpus-based techniques entail examining big datasets to spot biased linguistic patterns [21]. In the self-supervised machine learning method, bias is detected by human annotators who use labeled data to train machine learning models. The patterns found in the labeled data are used by the algorithms to learn how to categorize fresh occurrences of bias. Lexical, syntactic, and semantic details can all be features in these models. Convolutional neural networks (CNN) and recurrent neural networks (RNN) are examples of deep learning models that have been used, as well as support vector machines (SVM), random forests, and other methods [22]. The unsupervised machine learning techniques look for patterns or structures in the data that are concealed. When there are few labels on the data, they might be helpful for bias identification. Unsupervised bias detection methods include topic modeling (such as Latent Dirichlet Allocation) and clustering algorithms (such as k-means, and hierarchical clustering) [23]. The crowdsourcing strategy includes collecting annotations or judgments from human annotators to identify biased information. Annotators might be asked to assess the amount of bias in a given text or to indicate individual instances of biases. Bias can be identified by aggregating these judgments. Crowdsourcing tools such as Amazon Mechanical Turk have been utilized for this purpose [24]. The trim and fill method is a statistical method for addressing bias while accounting for inequalities. The trim and fill strategy entails finding missing or unpublished studies and calculating their possible impact on meta-analysis results. The steps in applying this method include estimating the number of missing studies, imputation of their impact sizes, and re-estimation of the total effect [25]. The Item response theory (IRT), a statistical framework for modeling item responses, is employed to disentangle bias from true group differences. They highlight the advantages of IRT in capturing the varying levels of item difficulty and discrimination across different scales of the MMPI (Minnesota Multiphasic Personality Inventory, a widely used psychological assessment tool), thus providing a more accurate assessment of the underlying constructs [26]. The feature importance analysis is used to find the possible bias sources. If there are hidden biases embedded in the data, they can be found by analyzing feature importance scores or by employing methods like permutation importance [27]. The Heuristic method manually evaluates the predictions provided by machine learning algorithms to identify any potential biases performed by human auditors or subject-matter experts. This approach can be time-consuming and resource-intensive but can provide valuable insights in some cases [28].

These techniques used for the detection of the biases are summarized in Table 1, as shown below based on various performance metrics (Accuracy, Precision, Recall, and F1-Score).

Table 1. Summarization of various techniques for the Detection of Biases

			Performance Metrics			
Ref. No.	Year	Techniques	Accuracy	Precision	Recall	F1-Score
[21]	2021	BERT Model	✓	×	×	×
[22]	2022	Self-Supervised Learning	✓	×	×	×
[23]	2020	Unsupervised Technique	✓	✓	✓	✓
[24]	2020	Crowdsourcing Strategy	×	×	×	×
[25]	2003	Trim and fill	×	×	×	×
[26]	2000	method	×	×	×	✓
[27]	2010	Item Response	✓	×	×	×
[28]	1991	Theory Feature Importance analysis Heuristic Method	×	×	×	×

As the various technique summarized in Table 1, we can easily conclude that the unsupervised technique is better than others as it evaluated all the performance metrics for the detection of biases.

These biases need to mitigate to reduce the overall impact of biases from the system. The process of the mitigation of biases is concerned with the avoidance of biases and minimizing their harmful consequences. The various techniques for the mitigation of biases are discussed in section 2.4.

2.2 Mitigation of Biases

Bias mitigation refers to the strategic and systematic efforts made to establish systems, processes, and environments that are more inclusive, equal, and unbiased. It recognizes that prejudices may have far-reaching consequences for people and society, and it tries to mitigate these consequences by actively addressing and minimizing biases across several domains. In recent times, several techniques such as GAN, DNNs, SMOTE, and many more have emerged to address the mitigation of biases. Here, we will provide an in-depth exploration of some significant mitigation techniques.

The GANs (Generative Adversarial Networks) strategy for mitigation is used to produce synthetic data that is identical to the original dataset but has less bias. The generative model is taught to generate representative samples, whereas the discriminative model learns to differentiate between actual and synthetic data [29]. The adversarial debiasing strategy involves training a model using an adversarial framework i.e., Deep Neural Networks (DNNs), where one component tries to predict the output while the other attempts to predict the protected characteristic. The model learns to produce accurate predictions while eliminating the dependency on protected qualities by concurrently optimizing both elements [30]. The ensemble approaches to reduce the bias by mixing predictions from numerous models trained on various subsets of the data. Ensemble models can deliver more balanced and fair results by pooling predictions [31]. The SMOTE (Synthetic Minority Over-sampling Technique) is used to apply to the minority class, increasing its representation and bringing it closer to the dominant class. This can lessen the prejudice brought on by the minority class's under-representation [32]. The effect of the majority class on the model's decision limits is decreased by deleting the majority class instances that create Tomek linkages (T-Links). This can help mitigate prejudices induced by the majority class's overrepresentation and give a more equitable representation of all classes [33]. The strategy of fairness-aware adversarial perturbation (FAAP), is also used to mitigate bias, by using feature selection and feature encoding. In Feature Selection, choosing features that are less likely to induce or magnify biases. In feature encoding, utilizing techniques that mitigate bias, such as removing direct identifiers or aggregating sensitive attributes to protect individual privacy [34]. The Regularization approach to inductive genetic programming (IGP), places extra restrictions on machine learning models during training to decrease biases. Techniques like L1 or L2 regularization penalize certain parameters, encouraging the model to focus on useful aspects and minimizing the influence of potentially biased qualities [35]. AI Fairness

360 (AIF360) is an open-source toolkit developed by IBM Research that includes a complete set of algorithms, measurements, and tutorials for identifying and eliminating bias in machine learning models. It offers tools for calculating bias measures, bias reduction techniques, and fairness visualization [36].

These techniques used for the mitigation of the biases are summarized in Table 2, as shown below based on various performance metrics (Accuracy, Precision, Recall, and F1-Score).

Table 2. Summarization of various techniques for the Detection of Biases

Ref. No.	Year	Techniques	Performance Metrics			
			Accuracy	Precision	Recall	F1-Score
[29]	2019	GAN Model	✓	×	×	×
[30]	2020	DNNs Model	✓	×	×	×
[31]	2022	Ensemble Learning Approach	×	×	×	✓
[32]	2020	SMOTE	×	✓	✓	×
[33]	2016	T-link	✓	✓	×	✓
[34]	2022	FAAP	✓	×	×	×
[35]	2001	Inductive Genetic Programming	✓	×	×	×
[36]	2020	AIF360	✓	×	×	×

As shown in Table 2, the summarization of various techniques for the mitigation of the biases. We can conclude that the SMOTE and T-link mitigation technique performs better than the other as it has evaluated various performance metrics.

After the mitigation of biases is performed then there is a need to check the degree of fairness of the system which is done by using the fairness metrics. Fairness metrics are a collection of measurements that let you spot the bias in your data or model. The different fairness metrics are discussed in detail in section 2.5.

2.3 Fairness Metrics

Fairness metrics are objective measures or statistical indicators used to quantify and analyze the fairness and justice of actions or results regarding various groups or persons [37]. These metrics give a formal and quantitative framework for analyzing and comparing the treatment of different demographic groups, considering the aspects such as gender, race, age, and socioeconomic status. In recent times, numerous fairness metrics have arisen to measure the degree of fairness. This section will delve into an extensive examination of fairness metrics performed after the mitigation of biases has been executed to check the degree of fairness of the system. Several fairness metrics are shown in Figure 2.

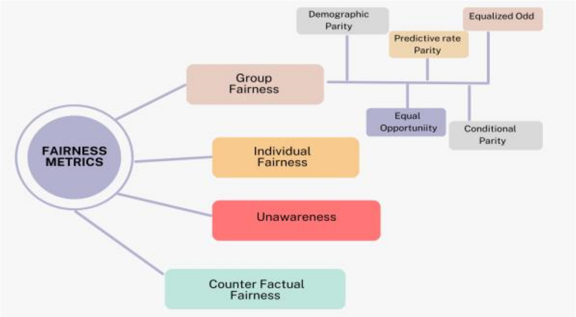


Fig. 2. Types of Fairness Metrics

In Group Fairness, equality must exist amongst various protected groups, such as those categorized by gender or race, for a fair conclusion to be possible. The various types of group fairness metrics [38], includes demographic parity, predictive rate parity, equalized odds, equal opportunity, and conditional parity. Demographic parity is also termed as independence, statistical parity, and disparate impact. This parity is attained when the likelihood of a certain prediction is not dependent on sensitive group membership. The requirement for demographic parity is, for all a, b $\epsilon\ A$ *then* $P\ (C = 1|A = a) = P\ (C = 1|A = b)$. The Predictive rate parity is obtained when the precision (or positive predictive values) in the subgroups is close to each other. Equalized odds are satisfied when the true positive rate (TPR) and (separately) the false positive rate (FPR) are the same across categories. A predictor z satisfies equalized odds concerning protected attribute x and outcome y, if: z and x are independent conditional on y i.e., P $(z{=}1\ |x = 0, y = a) = P\ (z = 1|x = 1, y = a), a\ \epsilon$ 0,1. The goal of equal opportunity is to achieve the same true positive rate across groups. A binary predictor z satisfies equal opportunity concerning protected attribute x and outcome y if: P $(z{=}1\ |x = 0, y = 1) = P\ (z = 1|x = 1, y = 1)$. And the Conditional statistical parity states that, given a set of legitimate factors L, Individuals in both the protected and unprotected categories should have an equal probability of receiving a favorable result [39]. For a set of legitimate factors l, predictor z, and protected attribute x satisfies conditional statistical parity if P (z $|l = 1, x = 0) = P\ (z\ |\ l = 1, x = 1)$. Similarly, individual fairness may be described as the treatment of persons who are comparable being those who are treated similarly [40]. The Unawareness, claim that the model is unaware of the sensitive qualities when they are purposefully removed from the data before the model is trained. The counter-factual fairness metric verifies whether a classifier yields the same result for one person as it does for another individual who is identical to the first except for one or more sensitive attributes [41].

These different fairness metrics used for the evaluation of biases are summarized in Table 3, as shown below based on various performance metrics (Accuracy, Precision, Recall, and F1-Score).

Table 3. Summarization of various techniques for the Fairness Metrics

Ref. No.	Year	Fairness Metrics	Performance Metrics			
			Accuracy	Precision	Recall	F1-Score
[37]	2021	Fairness Techniques	×	×	×	×
[38]	2021	Group Fairness	✓	×	×	×
[39]	2017	Conditional Parity	×	×	×	×
[40]	2018	Individual Fairness	✓	×	×	×
[41]	2017	Counter-factual Fairness	✓	×	×	×

As shown in Table 3, the summarization of various fairness metrics, we can conclude that the performance of group fairness, individual fairness, and counterfactual fairness is better than other fairness metrics as the accuracy is being evaluated in the case of these metrics.

3. Literature Survey

In this section, we offer a comprehensive overview of the existing surveys conducted on biases, detection of biases, mitigation of biases, and fairness metrics, along with a summary of their content. This examination allows us to identify the specific knowledge gap that our survey aims to fill. In this survey, to full fill the problem definition we have investigated 31 papers based on the tollgate approach [42]. The tollgate approach, which consists of five phases, facilitates the selection of 31 Primary Studies [43]-[47], [48]-[54], [55]-[61] and [62]-[73], in which we have summarized 5 papers for biases, 7 paper for the detection of biases, 7 paper for mitigation of biases and 12 papers related to fairness metrics in detail. This section is further divided into four subsections. Section 3.1 presents a research paper related to biases and section 3.2 presents a research paper related to the detection of biases. whereas section 3.3 presents a research paper related to the mitigation of biases and further section 3.4 presents a research paper related to fairness metrics.

3.1. Research Paper Related to Biases

On the topic of biases, Yahav et al. [43], discussed the examined comment mining using the term frequency-inverse document frequency (TF-IDF) technique and the inherent bias that might occur from word or phrase frequency differences between comment groups. The authors suggest re-weighting words' TF-IDF scores to counteract this bias and test their solution on Facebook fan page datasets. The bias reduction procedure yields more accurate and less biased findings. However, the paper's evaluation of large datasets and comparison to other bias detection and removal methods are limited. Speicher et al. [44], work on adult income datasets and work for various applications such as loan prediction, and criminal risk analysis which leads to algorithm bias due to using inequality indices that have been extensively studied in economics and social welfare. The limitation of this paper is that it works only with several attributes in combination. Xu et al. [45], propose an opinion mining model based on convolutional neural networks for enhancing recommendations (NeuO) which focuses on opinion bias, the two modules that build up NeuO are the MLP matrix factorization recommendation (MMF) module and the sentiment classification score (SC) module. For evaluating the proposed approach (NeuO) amazon dataset, the Yelp dataset, and the Taobao dataset is used. The main drawback associated with the suggested model is that the outcome of the suggested procedures is dependent on the review of the users. Lauw et.al. [46], aims to quantify the concepts of evaluation bias and disagreement within an evaluation system. This work proposes the Inverse Reinforcement or IR model and uses product reviews from the Epinions website. The main drawback of this paper is the proposed model needs to be verified using data from real-world situations. Feelders et al. [47], introduced a mixture modeling strategy to learning from data with selectivity bias, which is a common occurrence. This study demonstrated that a blend of two normal components can typically adequately reflect the distribution of financial ratios. This finding could be useful for other data mining applications in finance, such as bankruptcy prediction models. The major limitation of this work is the selection of the right parameter is important for improving the performance of the suggested model.

Table 4 below shows, the various publications based on the biases with the number of times the particular paper is referred for the publication and the venue of the publication.

Table 4. The list of publications for the biases along with the number of times the published paper referenced

Publications	Number of times Referenced	Venue
Yahav et al. [43]	84	TKDE
Speicher et al. [44]	251	SIGKDD
Xu et al. [45]	16	ICDM
Lauw et.al.[46]	42	TKDE
Feelders et al. [47]	6	KDD

3.2. Research Paper Related to Biases

In a study conducted by Norori et al. [48], bias is not commonplace in the medical industry but detecting and recognizing this bias can be difficult. Healthcare delivery is changing dramatically as an expanding number of data sources are shared, acquired, and incorporated into AI systems. AI is intended to help methods to clinical decision-making and public health efforts, eventually enhancing societal health. It is critical to include open scientific concepts in the creation and assessment of AI tools to create greater collaboration between the medical and AI disciplines and to provide a forum for varied opinions on the use of AI in medicine. Zhao et al. [49], presented a unique technique named LOGAN, which stands for Local Group Bias Detection Algorithm. LOGAN organizes instances based on their properties using a clustering approach, trying to maximize a biased measure (such as the performance gap between various groups) inside each cluster. This method, however, has a disadvantage in the form of potentially unequal cluster sizes. Furthermore, rather than directly analyzing protected properties for bias identification, which might be time-consuming, the architecture uses machine learning algorithms and performance measurements for each cluster independently. Zliobaite [50], studied and analyzed numerous strategies used to quantify discrimination in data and evaluate the efficacy of discrimination-aware prediction algorithms in a survey. The study emphasized the need of analyzing the fairness of prediction models routinely and objectively. The study's principal conclusion was that the majority of prior research has mostly focused on binary classification problems with binary-protected features. One weakness of this study is that it focuses primarily on statistical approaches for detecting discrimination in data. Kruse et al. [51], conducted an extensive literature review examining the challenges and opportunities that big data presents to the healthcare industry. The study highlighted the immense volume of healthcare data generated annually and emphasized the need for proper categorization and segregation of this data to ensure universal accessibility and transparency across healthcare institutions to reduce various biases that can occur in the system. The research investigations [52], [53], and [54] revealed a significant difficulty in the healthcare industry, namely the larger degree of unstructured data compared to other industries. Furthermore, this research emphasized that biased data might impair the capacity of healthcare decision-support systems to derive correct conclusions, thereby affecting performance and creating hazards to society.

Table 5 below shows, the various publications based on the detection of biases with the number of times the particular paper is referred for the publication and the venue of the publication.

Table 5. The list of publications for the detection of biases along with the number of times the published paper referenced

Publications	Number of times Referenced	Venue
Norori et al. [48]	77	Patterns
Zhao et al. [49]	12	arXiv
Zliobaite [50]	167	arXiv
Kruse et al. [51]	353	JMIR
Heudecker et al. [52]	26	Gartner
Chawla et al. [53]	526	JGIM
Jee et al.[54]	266	HIR

3.3. Research Paper Related to Mitigation of Biases

Akbari et al. [55], introduce the flatter loss, a unique loss function, that is used to mitigate bias in cross-dataset assessment of face age. In terms of decreasing bias and providing fairer forecasts across various demographic groupings, the suggested technique shows encouraging results. The results of this work add to the body of knowledge on bias mitigation in facial recognition tasks and emphasize the need for a standard method of bias mitigation for various applications to guarantee equity and inclusion. Hammond et al. [56], present a detailed overview of bias in medicine, emphasizing lessons learned from previous experiences and providing mitigation techniques. The authors emphasize the importance of increasing knowledge, education, and training to address biases in healthcare settings. They also emphasize the importance of technology, data collecting, and analysis in advancing equity and minimizing prejudice. This work is a great resource for healthcare professionals and policymakers, which demands the requirement of mitigating multiple biases in the system to build a fair prediction model. Mahabadi et al. [57], propose an end-to-end strategy for bias reduction in NLP models by explicitly modeling biases in corpora. The proposed system includes bias modeling and mitigation processes, as well as counterfactual instances and a unique loss function. The experimental results show promising outcomes in minimizing biases across several activities. The work contributes to the current research on bias reduction in NLP by emphasizing the significance of eliminating biases in training data to develop fair and unbiased language models. Bender et al. [58], present the idea of data statements to lessen system bias and advance scientific methods in NLP research. To promote openness, repeatability, and ethical concerns, the authors support the adoption of data declarations as industry best practices. In addition to highlighting the potential benefits of this practice in reducing biases and encouraging a better understanding and assessment of NLP models and systems, the study offers a structure and instructions for developing data statements. Wang et al. [59], discuss the problem of gender bias in deep image representations and highlights the drawbacks of reducing bias just by utilizing balanced datasets. The authors put forth a thorough methodology that comprises processes for data gathering, bias estimation, and bias avoidance. The experimental outcomes demonstrate how the suggested strategy effectively lowers gender bias and fosters fairness in deep learning models for image interpretation. The findings emphasize the significance of taking bias estimates and mitigation measures into account beyond dataset balance and add to the current research on bias reduction in computer vision. Maudaslay et al. [60], present a name-based counterfactual data replacement strategy to reduce gender bias in NLP models. The authors show how to successfully reduce gender bias in text creation tasks by substituting gendered pronouns with gender-neutral or underrepresented gender-associated pronouns. The results show the potential of counterfactual data replacement as a strategy for fostering equity and inclusion in language production systems, and they support continuing efforts to remove bias in NLP models. Hort et al. [61], present a counterfactual data replacement method based on names to lessen gender bias in NLP models. The authors demonstrate how replacing gendered pronouns with gender-neutral or underrepresented gender-associated pronouns can effectively eliminate gender bias in text production tasks. The findings demonstrate the potential of counterfactual data substitution as a tactic for promoting fairness and inclusion in language production systems and they complement ongoing initiatives to eliminate bias in NLP models.

Table 6 below shows, the various publications based on the mitigation of the biases with the number of times the particular paper is referred for the publication and the venue of the publication.

Table 6. The list of publications for the mitigation of biases along with the number of times the published paper referenced

Publications	Number of times Referenced	Venue
Akbari et al. [55]	13	ICPR
Hammond et al. [56]	16	JACC
Mahabadi et al. [57]	107	arXiv
Bender et al. [58]	612	TACL
Wang et al. [59]	292	ICCVW
Maudaslay et al. [60]	106	arXiv
Hort et al. [61]	28	ESEC

3.4. Research Paper Related to Fairness Metrics

On the topic of fairness Hyun et al. [62], presented a novel approach to data sanitization, data cleaning, and unfairness mitigation are all integrated into MLClean, a unified framework for data cleaning that facilitates the training of fair and accurate models. In addition to using the group fairness measures, the census income dataset and German credit dataset are also applied. The suggested method is used in several applications, including loan prediction and financial position. The main flaw in this MLClean is that it uses some specific steps of data cleaning namely data sanitization and cleaning and then unfairness mitigation, which is not able to clean the dataset up to the mark which leads to biases. Zhang et al. [63], the suggested method uses the causal graph methodology together with numerous path-specific effects and applies to a wide range of applications, including financial conditions, and income. The recommended technique applies the suggested measure using the adult income dataset, the Dutch Census dataset, and the group fairness metrics. The main flaw in this recommended strategy is that it requires the creation of discrimination-free prediction models. Hajian et al. [64], used four measures to develop a new pre-processing discrimination prevention methodology the direct discrimination prevention degree (DDPD), the direct discrimination protection preservation (DDPP), the indirect discrimination prevention degree (IDPD), and the indirect discrimination protection preservation (IDPP). The suggested approach utilized the adult income dataset and the German dataset and applies to applications like financial situations. It is necessary to research more discriminating measures to obtain more reliable results. Elbassuoni et.al. [65], present a measure for overcoming biases, namely the

Earth Mover's Distance between score distributions, using the dataset of simulated crowd workers. The suggested method applies to the online job market and group fairness is applied as a fairness metric. The problem with the proposed fairness metrics is that they require a different technique than Earth Mover's Distance. Kamishima et al. [66], present a new technique for the fairness-aware classifier Calders and Verwer's two-naive-Bayes (CV2NB), and it makes use of the adult dataset and Dutch Census dataset for that. The suggested technique uses an application, such as financial circumstances, and group fairness metrics are implemented. The suggested algorithm's time complexity must be improved, and that is its major weakness. Perez-Suay et al. [67], presented unique fair, and efficient nonlinear regression and dimensionality reduction approaches and added a component to the cost function based on the Hilbert-Schmidt independence criteria to ensure fair solutions and allow dealing with several sensitive variables at the same time. The suggested technique makes use of the adult dataset and the contraceptive method choice dataset, and it may be applied to a variety of applications such as loan prediction and medical condition prediction. The suggested framework's disadvantage is that it should be expanded to accommodate machine learning techniques. Kamishima et.al. [68], propose a regularization approach that applies to any prediction algorithm with probabilistic discriminative models. The suggested technique employs the German dataset and the Adult Income dataset, and it may be used for applications like financial position. The recommended technique is subject to group fairness, and the suggested strategy must be put into practice. Mancuhan et al. [69], provided a new training set correction method for its SVM classification system that uses discrimination prevention. The suggested method makes use of the group fairness measures and the German credit dataset. The recommended approach can be used in a variety of contexts, including credit. The main problem is that new classification methods must be created without the use of rectified training data. Luong et al. [70], modeled the discrimination discovery and prevention problems by a variant of k-NN classification that implements the legal methodology of situation testing for that it uses accuracy measure. The suggested method is employed with the German credit dataset, adult dataset, census dataset, and criminal dataset to address a variety of scenarios including criminal risk, financial status, and credit. This work makes use of individual fairness metrics. The main flaw in the suggested strategy is that it varies depending on the decision attributes (an attribute recording the historical decisions) very much. Kamishima et al. [71], suggested a fresh approach, an actual fair-factorization technique, and showed that it significantly enhanced performance. This technique employs Calders and Verwer's score (CVS) and normalized prejudice index (mutual information). The adult income dataset and the Dutch census dataset are employed, and the proposed approach is relevant to financial situations, loans, and so on. The proposed major is subjected to the group fairness metrics. The key disadvantage is that the proposed approach must be extended to increase performance. Zliobaite et al. [72], proposed the discrimination-aware classification paradigm in the presence of explanatory qualities that relate to the sensitive attribute, for which it employs a grouping approach. The Dutch census dataset and adult income dataset are utilized, and the suggested technique is applied to various aspects such as financial status and income. The group fairness metrics are applied to the technique, and the suggested approach functions well only when the sensitive characteristic and the explanatory attribute have a significant association. Kamiran et al. [73], demonstrate and evaluate two novel discrimination-aware classification methods. These simple and adaptable techniques enable traditional probabilistic classifiers (ROC) and classifier ensembles (DAE) discrimination-aware by using decision theory. The datasets on communities and crimes and adult income are utilized and can be applied to a variety of applications such as criminal risk and financial status. The disadvantage of the proposed technique is that it must be adjusted to handle various circumstances.

Table 7 below shows, the various publications based on the fairness metrics with the number of times the particular paper is referred for the publication and the venue of the publication.

Table 7. The list of publications for the fairness metrics along with the number of times the published paper referenced

Publications	Number of times Referenced	Venue
Hyun et al. [62]	48	DEEM
Zhang et al. [63]	33	TKDE
Hajian et al. [64]	368	TKDE
Elbassuoni et.al. [65]	12	EDBT
Kamishima et al. [66]	18	DMKD
Perez-Suay et al. [67]	82	PKDD
Kamishima et.al. [68]	405	ICDMW
Mancuhan et al. [69]	9	ICDMW
Luong et al. [70]	202	SIGKDD

Kamishima et al. [71]	16	ICDMW
Zliobaite et al. [72]	23	ICDM
Kamiran et al. [73]	311	ICDM

Discussion

After going through the survey in section 3, based on biases, detection of biases, mitigation of biases, and fairness metrics we have come across various flaws which are summarized in this section. Research needs to be conducted to address the existing shortcomings to enhance the system's performance. The major limitation regarding the biases in the system is the necessity to minimize the biases when the dataset is inadequate [43]. According to the study [44], working with a protected attribute that lacks proper representation, the model's performance is impeded, and more effort is required. Therefore, it is crucial to ensure a well-represented protected attribute. In the case of the detection of biases the major loophole is stated in [49], according to this work rather than directly analyzing the protected attribute (which is the major reason for biases) for bias detection, they worked on the whole dataset which leads to increase in time complexity, computational costs and also degrades the overall performance of the system. The mitigation of biases also deals with several drawbacks, one of the significant limitations is described in [55], according to this study, there is a necessity for identifying the various method to carry out the mitigation of biases among various stages of the model. In the paper [56], the author depicted the need for a strategy for the mitigation of multiple biases in the dataset to build a fair prediction model. In the case of fairness metrics, the major challenges are described in [64], according to this study, it is necessary to find out more discrimination or fairness measures to obtain more reliable results. By seeing all these flaws, we derived certain challenges which are summarized below.

Challenges

In this section, we identify and discuss six major challenges drawn from the collected papers, emphasizing the need for more study or expansions of current work. These difficulties perform as a wake-up call, highlighting regions that need more investigation.

- Minimizing biases when working with inadequate datasets.
- Ensuring proper representation of protected attributes.
- Developing efficient and direct methods for bias detection in datasets.
- Identifying the effective approaches for mitigating biases at various stages of the model.
- Developing strategies to effectively mitigate multiple biases in the dataset to build a fair prediction model.
- Exploring and refining fairness metrics to achieve more dependable outcomes.

Research Questions

To overcome the above challenges, we have framed the research questions based on biases and fairness in the dataset for enhancing the performance of the model which are as follows.

RQ 1. What are the different ways to minimize biases when working with inadequate datasets?

RQ 2. What are the different ways to identify the well representation of the protected attributes effectively?

RQ 3. What are the efficient and direct methods for bias detection in the dataset?

RQ 4. What are the effective methods for mitigating biases at various stages of the model?

RQ 5. What strategies can be developed to effectively mitigate multiple biases in the dataset to build a fair prediction model?

RQ 6. What are the fairness measures to be explored to obtain more reliable results?

4. Findings

In the literature, there are many definitions of biases, detection of biases, mitigation of biases, and fairness approaches but there are many opportunities still available for research work related to all. In this section, we provided the solutions to the research questions.

RQ 1. What are the different ways to minimize biases when working with inadequate datasets?

Data bias occurs when the source data is skewed, providing results that are not fully representative of the target population. This problem of representing of dataset accurately with the target population arises due to insufficient data i.e. when the dataset cannot accurately reflect the target population because there is not enough data to represent the whole population. This condition leads to data biases.

Nowadays, data must be unbiased for the better performance of any model. For the solution to the above-mentioned research question, there is a need for proper data standardization. This will help to make data more effective and interoperable.

If data standardization is not followed such as the International Organization for Standardization (ISO) including data quality and management (ISO 8000 series), data exchange and interchange (ISO 20022), metadata management (ISO 11179), and many more [74]. As in the case of data quality and management (ISO 8000 series) the key standards are:

- ISO 8000-1:2017 - Data Quality Framework: This standard defines the overall foundation for data quality management. It includes ideas, principles, and terminology relevant to data quality, as well as helps in planning, implementing, and measuring data quality inside a dataset.
- ISO 8000-2:2017 - Data Quality Model: This is concerned with the creation of a data quality model. It explains how to build a model that describes the qualities of high-quality data, such as correctness, completeness, consistency, timeliness, and relevance.
- ISO 8000-3:2017 - Data Quality Measurement: This standard concerns the measuring of data quality. It assists in designing data quality measures, establishing measurement methodologies, and evaluating data quality outcomes to analyze and monitor data quality.
- ISO 8000-4:2016 - Data Quality Assessment: ISO 8000-4 gives instructions on how to perform data quality assessments. Although the standard does not expressly address datasets, organizations may modify the evaluation processes and approach to evaluate the quality of their datasets. It aids in the identification of gaps and opportunities for improvement in the dataset's quality.
- ISO 8000-5:2021 - Data Quality Planning: This standard emphasizes the need for data quality planning. It includes assistance in identifying data quality goals, creating quality criteria, implementing quality controls, and building a data quality management strategy inside a dataset.
- ISO 8000-6:2018 - Data Quality Management: ISO 8000-6 defines the method to manage data quality throughout its lifespan. It discusses data quality roles and duties, data quality policies and procedures, and data quality management integration for datasets.

If the data standard is not followed, it becomes challenging to interoperate, analyze, and understand the data which further leads to data biases. Also, by including proper data points (discrete units of information), data bias can be reduced. More comprehensive and proper data can help the algorithms to minimize the data biases within the dataset. Next to overcome the problem of insufficient data several techniques can be applied, such as data resampling, data augmentation, and data gathering.

RQ 2. What are the different ways to identify the well representation of the protected attributes effectively?

A "well representation" of a protected attribute means that the dataset is accurately representing the diversity of the population concerning that attribute. Also, the well representation of the protected attribute in the dataset indicates that the trained model that employs the dataset learns appropriate patterns about the group. If the protected attribute is not adequately represented, it can lead to bias issues. We have conducted statistical testing to ensure that the protected characteristic is well represented, for that, we have gone through various statistical testing such as the z-test, Chi-square test, one-way ANOVA test, and two-way ANOVA. These statistical tests are used to identify the well representation of protected attributes efficiently.

• Z-test: The z-test is used to compare the means of protected attributes across different groups, such as gender or race. If the means are significantly different, this can indicate that the protected attribute is not well represented which leads to a potential bias in the data. For the evaluation of z-value, we use the following formula:

$$Z = (X - \mu) / \sigma$$

Where X represents the raw data,

μ *is the mean of the population and*

σ *is the standard deviation for the population.*

- Chi-square test: The Chi-square test is used to test the association between protected attributes and other variables. If there is a significant association, this can indicate a potential bias or discrimination. The formula for chi-square is:

$$\chi^2 = (O_i - E_i)^2 / E_i$$

Where O_i *is observed value* (*actual value*),

E_i *is expected value.*

- One-way ANOVA: Compared to z-test, the one-way ANOVA test is useful when the protected attribute has more than two categories. The one-way ANOVA test is used to compare the means of protected attributes across different groups, such as age or disability status. The formula for one-way ANOVA is:

$$F = MSB/MSE$$

Where F is F-statistic,

MSB is the Mean squares between groups,

MSE is the Mean squares of errors.

- Two-way ANOVA: The two-way analysis of variance (ANOVA) is used to investigate the relationship between protected attributes and other variables, such as job performance or salary. If there is a significant interaction, this can indicate a potential bias or discrimination. In the case of two-way ANOVA, the F-statistic for each source of variation can be calculated as:

$$F = MS_A/MSE$$

$$F = MS_B/MSE$$

Where F is F-statistic,

MS$_A$ *is the Mean square corresponding to group A,* MS$_B$ *is the Mean square corresponding to group B, and MSE is the Mean square Error.*

For instance, we have analyzed the cardiovascular disease dataset (https://www.kaggle.com/datasets/sulianova/cardiovascular-diseasedataset) that has 70000 data points of which 24470 male and 45530 female. We are interested in determining whether gender is a well-represented protected attribute in the dataset or not.

For this, we performed a z-test, and we calculated the expected proportion of females based on the data points. Let's say that proportion is 0.65. We would then calculate the standard error of the proportion which is:

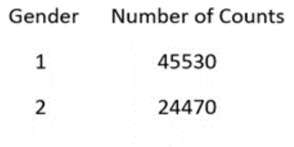
$\sqrt{(0.65 * 0.65/100)}$ = 0.65 then, calculated the z-score:

$$z = (0.34 - 0.65) / 0.065 = -4.76$$

Finally, we compared the z-score to a critical value based on the desired level of significance (e.g., 0.05). For a two-tailed test, the critical value would be approximately +/- 1.96. Since the calculated z-score (-4.76) is less than the critical value (-1.96), we conclude that there is a statistically significant difference between the proportion of females and males in the dataset. This would suggest that gender is not a well-represented protected attribute in the cardiovascular disease dataset.

RQ 3. What are the efficient and direct methods for bias detection in the dataset?

For efficient and direct detection of bias in the dataset, we have proposed a novel methodology using the MapReduce framework with a class imbalance approach which directly categorizes protected attributes rather than working on the whole dataset. In this methodology, we have three modules as Clustering module, the MapReduce framework module, and the class imbalance module. The cardiovascular disease dataset (https://www.kaggle.com/datasets/sulianova/cardiovascular-diseasedataset) has been given as input (where gender is the protected attribute) to the first module, further, the dataset is split into the form of clusters in the clustering module with the help of a clustering algorithm (K-Means clustering). Thereafter, the MapReduce strategy is applied to the cluster data sets within the second module i.e., the MapReduce Framework module to categorize the protected attribute, in which the mapper maps the clustered input data into ⟨Key, value⟩ pairs then these intermediate key-value pairs are shuffled and then sends to the reducer. Then, the reducer reduces the input data in ⟨Key, Total values⟩ pairs which count the overall sum of keys. The output of the MapReduce framework gives the overall counts of the protected attributes i.e., gender (where 1 represents female and 2 represents male). Figure 3, below shows the output of the MapReduce framework.



Gender	Number of Counts
1	45530
2	24470

Fig. 3. The Output of MapReduce Framework regarding Protected Attribute i.e., gender

This output is sent to the third module which is the class imbalance module. In the class imbalance module, the disorder is tested on the categorized protected attribute by using the Shannon entropy. The Shannon entropy varies between 0 to 1 and tends to 0 in the case of the disorder dataset, in this case, the value of **Shannon entropy is 0.28** which means disorder exists in the dataset. After identifying the disorder then bias is detected by applying the class imbalance approach which is the balance formula, which gives the percentage of bias existing in the dataset. In this case, the value of the balance formula is 0.28 means 28% bias exits in the dataset, which is gender bias. Table 8, shows the output of disorder in the dataset for cardiovascular disease dataset.

Table 8. Bias Detection Values for Gender

Disorder test	Values
Shannon Entropy	0.28

Table 9, represents the bias detection values for gender in the case of cardiovascular disease dataset.

Table 9. Bias Detection Values for Gender

Class Imbalance Approach	Values
Balance Formula	0.28

Hence, the proposed framework reduces the time complexity as well as the overall computational costs of the system.

RQ 4. What are the effective methods for mitigating biases at various stages of the model?

Mitigating data biases is the process of removing biases from data. To overcome favouritism, data bias must be reduced. If the data is skewed, the model's performance also suffers. Systems become more reliable and robust after the removal of biases from data. To follow the standardized approach for bias mitigation, the system might implement several bias mitigation strategies. The main goal is to increase model accuracy while making sure that the models are less biased toward sensitive or protected features. In DSS, there are three stages where biases can occur such as pre-processing, in-processing, and post-processing. To mitigate these biases, we can apply different strategies in the above three stages which are shown in Figure 3.

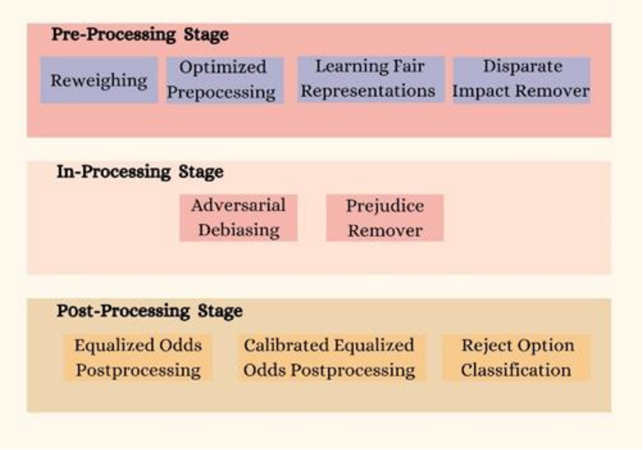


Fig. 4. Bias Mitigation strategies

The pre-processing stage is used to mitigate biases present in the training data. There are four bias mitigation strategies that we have applied in pre-processing stage including Reweighing, Optimized Pre-processing, Learning Fair Representations, and Disparate Impact remover. Reweighting is a bias mitigation approach that is often used when the distribution of sensitive attributes within a dataset is unbalanced. Reweighting adjusts sample weights (sensitive attribute values) to reflect the real population distribution. Each sample is weighted by the ratio of the real population frequency of the sensitive attributes to the dataset frequency. To maintain dataset weight, each sample's weight is multiplied by a scaling factor. The next bias mitigation technique is optimized pre-processing, which is used when there is a known or suspected source of bias in the input data, such as gender, race, or age. Optimized pre-processing works by pre-processing the input data in a way that reduces or eliminates the impact of biased variables on the models. The succeeding bias mitigation strategy is learning fair representation, which is a demographic bias (certain groups are under-represented or unfairly represented compared to others) mitigation technique. In general, the work of learning fair representation is to learn a representation of the input data that is both informative for the task at hand and fair concerning protected attributes. The last strategy in the pre-processing stage is disparate impact remover employed when training data represents a protected group. Disparate impact remover alters the model decision boundary to produce equivalent results for all protected groups. In the next stage which is the in-processing stage, we have two bias mitigation strategies adversarial debiasing and prejudice remover. Adversarial debiasing is used when a dataset contains bias that could negatively impact decision-making. The adversarial debiasing works by training the model to minimize the correlation between the predictions and the sensitive attribute. The next in-processing strategy is prejudice remover, the idea here is to add an optimization technique that is discrimination-aware in the training objective, for this purpose training of the dataset should be done properly. The last stage which is the post-processing stage for bias mitigation contains three strategies equalized odds postprocessing, calibrated equalized odds postprocessing, and reject option classification. Equalized odds postprocessing involves adjusting the model's predictions after they have been made to ensure that the false positive and false negative rates are equal across different groups defined by a sensitive attribute. In equalized odd the weights (value of sensitive attributes) are applied to the predicted probabilities for each group to adjust the predictions and ensure that the false positive rates (FPR) and false negative rates (FNR) are equalized. The calibrated equalized odds postprocessing is a variant of equalized odds postprocessing that considers the validation of the model's predicted probabilities as well as the false positive and false negative rates. This involves using a set of weights to adjust the model's predicted probabilities for each group to ensure that the FPR and FNR are equalized across groups, while also maintaining the calibration of the predicted probabilities. The last strategy of the post-processing stage is rejecting option classification which involves adding a reject option to the model's predictions, that allows the model to abstain from making a prediction when the model is uncertain about the correct label. The reject option classification applies the reject option to the model's predictions and defines a threshold for the confidence score (a measure of a model's predictability for a certain case) of the model's predictions and rejects instances that fall below this threshold.

These are the numerous strategies to mitigate biases in different stages of the model. Biases will be reduced by using these techniques in various situations.

RQ 5. What strategies can be developed to effectively mitigate multiple biases in the dataset to build a fair prediction model?

It is necessary to develop an effective way to mitigate multiple biases in the system to build a fair prediction model to improve the performance i.e., if the proposed model correctly predicts the outcome, this will improve model accuracy, which will improve performance.

For the solution to this research question, there is a need for a robust data-cleaning process to improve fairness within the prediction model. This will minimize the impact of the biases in the system and the system become fairer. That is why we have implemented a data-cleaning pipeline that can mitigate multiple biases that occur in the cardiovascular disease dataset. The pipeline comprises various modules for data cleaning such as the deduplication of data module, irrelevant data removal module, fixing structural errors module, filtering outliers module, handling missing data module, and validating the data. The data cleaning pipeline with the various module is shown in Figure 5.

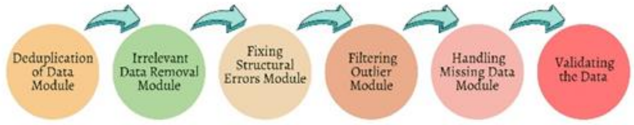

Deduplication of Data Module
Irrelevant Data Removal Module
Fixing Structural Errors Module
Filtering Outlier Module
Handling Missing Data Module
Validating the Data

Fig. 5. Data Cleaning Pipeline

In the deduplication of the data module, the goal is to detect and eliminate duplicate data to enhance data quality and guarantee correct analysis and decision-making. To implement the deduplication of data various techniques are applied such as Query-based process, ETL (Extract, Transform, Load) process, and File-based process. The query-based process typically involves comparing the data in each record with the data in other records to determine if there are any matches. For this purpose, the query-based process works by first defining a set of query criteria that will be used to identify potential duplicate records. Once the query criteria have been defined, the process involves running a query against the dataset to identify potential duplicates. The ETL process is applied for deduplication by taking data from many sources, changing it to a standard format, and then loading it into a single place where duplicates can be found and deleted. A file-based process includes comparing files to detect and eliminate duplicate data copies. This method is often used in backup and recovery systems, where several backups of the same data may be generated over time. This was accomplished via the use of numerous software tools and platforms such as Commvault, veritas netbackup, Microsoft Windows server deduplication, etc. This various deduplication is applied to data and the output is sent to the succeeding module. The second module is the irrelevant data removal module which will remove the insignificant data. To implement the removal of this data different machine learning algorithms are applied such as the K-Means algorithm and then the output is sent to the next module. The third module is the fixing of structural errors which will remove the errors that occur during data measurement, data transmission, or other related processes. Various procedures are used for fixing structural errors, including completing missing fields and ratings and fixing pages with manual actions. Sometimes testing tools are also used for this purpose, and the output is passed to the next module. The fourth module is the filtering outlier module which will remove the extreme values that significantly deviate from the dataset value. To perform the outlier removal various methods are used such as Box plots, IQR method, Z-Score method, and Distance from the mean method, and the result is sent to the last module. The fifth module will handle the missing data module which will manage the part of the observations in a data collection that are empty. To handle missing data various approaches can be accomplished such as finding the most probable value (through ML algorithm), filling in the missing values manually, and ignoring them missing values. After passing through all these modules of the data cleaning pipeline, multiple biases will be eliminated from the dataset to build a fair prediction model as disparity from the system is removed which is the main cause of biases in the system. Further, to verify the fair prediction model the validation of the data is performed in which verification of the quality and correctness (i.e., checking the remaining error, to remove from the dataset) of the source data is done. For validation, various approaches are used such as data validation functions (Range check, Type check, Check digit), validating data with a data processor, and data validation through API.

Moreover, by keeping a higher degree of planning and awareness throughout the system's development, many common mistakes and issues may be avoided. However, putting such ideas into practice poses new data management research issues.

RQ 6. What are the fairness measures to be explored to obtain more reliable results?

To obtain more reliable results we need to remove multiple biases that can arise in the system, therefore it is necessary to explore the hybrid fairness metrics model. Based on the literature survey, we have seen that maximum biases occur due to disparity in the dataset. So, the disparity is a major concern for the fairness of the system. Secondly, along with the dataset we have to check the model on which the dataset is trained to be fair, and at last, we need to validate the error of the system so, that we can obtain reliable results.

To overcome all these issues, we have implemented the hybrid fairness metrics for that we have used the cardiovascular disease dataset. The hybrid fairness metrics include three stages namely, Conditional Demographic Disparity, Fairness-Aware Regularization, and Fairness in Error. The pictorial representation of all three stages is given in Figure 6.

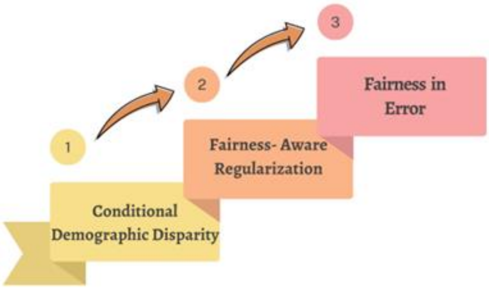

1
2
3
Conditional Demographic Disparity
Fairness- Aware Regularization
Fairness in Error

Fig. 6. The Hybrid Fairness Metrics

The stage first is conditional demographic disparity, which evaluates disparities in predictive outcomes across different demographic groups while considering additional relevant features. It focuses on determining fairness within specified subgroups based on different features.

The formula for Conditional demographic disparity:

$$CDD = \left({}^{1}/_{n}\right) * \Sigma_i \ n_i * \ DD_i$$

Where

$\Sigma_i \ n_i$ = n is the total number of observations,

n_i is the number of observations for each subgroup,

$DD_i = n_i^{(0)}/n(0) - n_i^{(1)}/n\binom{11}{1} = \ P_i^R(y^0) - P_i^A(y^1)$ is the demographic disparity for the ith subgroup.

The stage second is fairness-aware regularization, which incorporates fairness constraints into the model training process. It penalizes incorrect predictions and promotes the model to make correct judgments. It may be seen as a regularization word that balances accuracy and fairness aims.

Fairness-Aware Regularization can be formulated as:

Loss with regularization(θ) = *Loss*(θ) + $\lambda * R(\theta)$

Where

θ *represents the parameters of the model.* Loss(θ) *is the original loss function used for model training.* λ is the regularization strength hyperparameter that controls the trade-off between fairness and accuracy whereas R(θ) *is the regularization term for fairness.*

The last stage which is fairness in error, compares the error rates of various groups. It investigates whether the model's errors are spread proportionately among groups, representing the fairness of the error rates.

The formula for fairness in error includes such as to Identify the sensitive attribute or demographic group of interest, denoted as G. Define the error rate for each group. Let's denote the error rate for group G as Err(G). Calculate the overall error rate of the model, denoted as Err (Overall).

Compute the fairness in error metric by comparing the error rates between the groups:

Fairness in Error = |*Err* (*G*) – *Err* (*Overall*)|

The fairness in error metric represents the absolute difference between the error rate of the specific group and the overall error rate. The formula quantifies the discrepancy in error rates between the group of interest and the overall error rate. A smaller value of *Fairness in Error* indicates a more equitable distribution of errors across groups.

After implementing the hybrid fairness metrics, we applied the cardiovascular disease dataset to evaluate the degree of fairness of the system and found that it performs better than the heuristic fairness metrics for evaluating the degree of fairness in the system.

5. Conclusion and Future Work

In this study, we provide a broad overview of the biases, detection of biases, mitigation of biases, and fairness metrics that appear in decision support systems. We have performed this review after critically analyzing 31 relevant research papers published in a well-known publication. After reviewing we identify certain challenges which are needed to be full fill for better performance of DSS and based on that we framed the research questions such as minimizing biases when working with inadequate datasets, ensuring well representation of protected attributes, developing efficient and direct methods for bias detection, identifying the effective approaches for mitigating biases at various stages of the model, developing strategies to effectively mitigate multiple biases in the system to build a fair prediction model and exploring and at last, refining fairness metrics to achieve more dependable outcomes. We have also provided the solutions to the depicted research questions. Furthermore, simply having a greater level of awareness and forethought throughout the data gathering might prevent frequent mistakes and problems. It is hoped that through broadening the readers' horizons, they would be inspired to think critically as they develop systems or methodologies that are less likely to be harmful to or biased towards a specific group. Researchers need to take this issue seriously and broaden their understanding in this area as decision-support systems become more prevalent. Additional future endeavors and directions include the automated detection of the biases such as data bias, algorithmic bias, and human bias that can occur in the decision support system, robust fair predictions and mitigation of biases, and how to clean semi-structured and unstructured data rather than only structured data so that biases can be minimized up to the mark.

Acknowledgments

We extend our deepest appreciation to Dr. Md. Tanwir Uddin Haider for his profound dedication, relentless encouragement, and invaluable guidance. Without his exceptional support, this research would not have reached its full potential. We are forever grateful for his remarkable contributions, which have undoubtedly left an indelible mark on this study and our academic growth.

Author contributions

Jyoti Prakhar: Reviewed Literature Survey, Framed Research Questions with Solutions, Writing an original draft.

Dr. Md. Tanwir Uddin Haider: Discussed the literature survey, and Identified Challenges.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Power DJ. Understanding data-driven decision support systems. Information Systems Management. 2008 Mar 28;25(2):149-54.

[2] Dastin J. Amazon scraps secret AI recruiting tool that showed bias against women. In Ethics of data and analytics 2018 Oct 10 (pp. 296-299). Auerbach Publications.

[3] Holstein T, Dodig-Crnkovic G. Avoiding the intrinsic unfairness of the trolley problem. In Proceedings of the International Workshop on Software Fairness 2018 May 29 (pp. 32-37).

[4] Power DJ. Decision support systems: concepts and resources for managers. Greenwood Publishing Group; 2002.

[5] Power D. What are the features of a data-driven DSS? DSS News. 2007 Feb;8(4):6.

[6] Balayn A, Lofi C, Houben GJ. 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. 2021 Sep;30(5):739-68.

[7] Srinivasan R, Chander A. Biases in AI systems. Communications of the ACM. 2021 Jul 26;64(8):44-9.

[8] Torralba A, Efros AA. An unbiased look at dataset bias. InCVPR 2011 2011 Jun 20 (pp. 1521-1528). IEEE.

[9] Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability, and transparency 2018 Jan 21 (pp. 77-91). PMLR.

[10] Srinivasan R, Chander A. Crowdsourcing in the Absence of Ground Truth–A Case Study. arXiv preprint arXiv:1906.07254. 2019 Jun 17.

[11] Plous S. The psychology of judgment and decision making. Mcgraw-Hill Book Company; 1993.

[12] Kahneman D. Evaluation by moments: Past and future. Choices, values, and frames. 2000 Sep 25:693-708.

[13] Hao K. This is how AI bias happens—and why It’s so hard to fix, 2019.

[14] Elwert F, Winship C. Endogenous selection bias: The problem of conditioning on a collider variable. Annual review of sociology. 2014 Jul 30;40:31-53.

[15] Quinn T. Judea Pearl and Dana Mackenzie: THE BOOK OF WHY: The new science of cause and effect. TLS. Times Literary Supplement. 2018 Sep 21(6025):31-2.

[16] Kilbertus N, Ball PJ, Kusner MJ, Weller A, Silva R. The sensitivity of counterfactual fairness to unmeasured confounding. In Uncertainty in artificial intelligence 2020 Aug 6 (pp. 616-626). PMLR.

[17] Friedman B, Nissenbaum H. Bias in computer systems. ACM Transactions on Information Systems (TOIS). 1996 Jul 1;14(3):330-47.

[18] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR). 2021 Jul 13;54(6):1-35.

[19] Amatriain X. What does the concept of presentation-feedback bias refer to in the context of machine learning? quora, 2015.

[20] Austin PC, Platt RW. Survivor treatment bias, treatment selection bias, and propensity scores in observational research. Journal of clinical epidemiology. 2010 Feb 1;63(2):136-8.

[21] Schick, Timo, Sahana Udupa, and Hinrich Schu¨tze. ”Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP.” Transactions of the Association for Computational Linguistics 9 (2021): 1408-1424.

[22] Sirotkin, Kirill, Pablo Carballeira, and Marcos Escudero-Vin˜olo. ”A study on the distribution of social biases in self-supervised learning visual models.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[23] Field, Anjalie, and Yulia Tsvetkov. ”Unsupervised discovery of implicit gender bias.” arXiv preprint arXiv:2004.08361 (2020).

[24] Lim, Sora, et al. ”Annotating and analyzing biased sentences in news articles using crowdsourcing.” Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020.

[25] Terrin, Norma, et al. ”Adjusting for publication bias in the presence of heterogeneity.” Statistics in medicine 22.13 (2003): 2113-2126.

[26] Waller, Niels G., Jane S. Thompson, and Ernst Wenk. ”Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales: an illustration with the MMPI.” Psychological Methods 5.1 (2000): 125.

[27] Altmann, Andr´e, et al. ”Permutation importance: a corrected feature importance measure.” Bioinformatics 26.10 (2010): 1340-1347.

[28] Smith, James F., and Thomas Kida. ”Heuristics and biases: Expertise and task realism in auditing.” Psychological bulletin 109.3 (1991): 472.

[29] Abusitta, Adel, Esma A¨ımeur, and Omar Abdel Wahab. ”Generative adversarial networks for mitigating biases in machine learning systems.” arXiv preprint arXiv:1905.09972 (2019).

[30] Darlow, Luke, Stanis law Jastrzebski, and Amos Storkey. ”Latent adversarial debiasing: Mitigating collider bias in deep neural networks.” arXiv preprint arXiv:2011.11486 (2020).

[31] Nascimento, Francimaria RS, George DC Cavalcanti, and Marjory Da CostaAbreu. ”Unintended bias evaluation: An analysis of hate speech detection and gender bias mitigation on social media using ensemble learning.” Expert Systems with Applications 201 (2022): 117032.

[32] Tarawneh, Ahmad S., et al. ”Smotefuna: Synthetic minority over-sampling technique based on furthest neighbour algorithm.” IEEE Access 8 (2020): 59069-59082.

[33] Elhassan, T., and M. Aljurf. ”Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method.” Global J Technol Optim S 1 (2016): 2016.

[34] Wang, Zhibo, et al. ”Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[35] Nikolaev, Nikolay Y., and Hitoshi Iba. ”Regularization approach to inductive genetic programming.” IEEE Transactions on evolutionary computation 5.4 (2001): 359-375.

[36] Hufthammer, Knut T., et al. ”Bias mitigation with AIF360: A comparative study.” Norsk IKT-konferanse for forskning og utdanning. No. 1. 2020.

[37] Zhou N, Zhang Z, Nair VN, Singhal H, Chen J, Sudjianto A. Bias, Fairness, and Accountability with AI and ML Algorithms. arXiv preprint arXiv:2105.06558. 2021 May 13.

[38] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR). 2021 Jul 13;54(6):1-35.

[39] Ritov YA, Sun Y, Zhao R. On conditional parity as a notion of non-discrimination in machine learning. arXiv preprint arXiv:1706.08519. 2017 Jun 26.

[40] Verma S, Rubin J. Fairness definitions explained. In Proceedings of the international workshop on software fairness 2018 May 29 (pp. 1-7).

[41] Russell C, Kusner MJ, Loftus J, Silva R. When worlds collide: integrating different counterfactual assumptions in fairness. Advances in neural information processing systems. 2017;30.

[42] Afzal W, Torkar R, Feldt R. A systematic review of search-based testing for non-functional system properties. Information and Software Technology. 2009 Jun 1;51(6):957-76.

[43] Yahav I, Shehory O, Schwartz D. Comments mining with TF-IDF: the inherent bias and its removal. IEEE Transactions on Knowledge and Data Engineering. 2018 May 24;31(3):437-50.

[44] Speicher T, Heidari H, Grgic-Hlaca N, Gummadi KP, Singla A, Weller A, Zafar MB. A unified approach to quantifying algorithmic unfairness: Measuring individual group unfairness via inequality indices. In Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery data mining 2018 Jul 19 (pp. 2239-2248).

[45] Xu Y, Yang Y, Han J, Wang E, Zhuang F, Xiong H. Exploiting the sentimental bias between ratings and reviews for enhancing recommendation. In2018 IEEE international conference on data mining (ICDM) 2018 Nov 17 (pp. 1356-1361). IEEE.

[46] Lauw HW, Lim EP, Wang K. Bias and controversy in evaluation systems. IEEE Transactions on Knowledge and Data Engineering. 2008 Apr 25;20(11):1490-504

[47] Feelders AJ, Chang S, McLachlan GJ. Mining in the Presence of Selectivity Bias and its Application to Reject Inference. InKDD 1998 Aug 27 (pp. 199-203).

[48] Norori, Natalia, et al. “Addressing bias in big data and AI for healthcare: A call for open science.” Patterns 2.10 (2021): 100347.

[49] Zhao, Jieyu, and Kai-Wei Chang. “LOGAN: Local group bias detection by clustering.” arXiv preprint arXiv:2010.02867 (2020)

[50] Zliobaite, Indre. “A survey on measuring indirect discrimination in machine learning.” arXiv preprint arXiv:1511.00148 (2015).

[51] Kruse, Clemens Scott, et al. ”Challenges and opportunities of big data in health care: a systematic review.” JMIR medical informatics 4.4 (2016): e5359.

[52] Heudecker N. “Hype Cycle for Big Data.” Gartner. URL: https://www.gartner.com/doc/2574616/ hype-cycle-big-data- [accessed 2016-11-08] [WebCite Cache ID 6lsI6Sxxr] 2013 Jul 31.

[53] Chawla, Nitesh V., and Darcy A. Davis. “Bringing big data to personalized healthcare: a patient-centered framework.” Journal of general internal medicine 28.3 (2013): 660-665.

[54] Jee, Kyoungyoung, and Gang-Hoon Kim. “Potentiality of big data in the medical sector: focus on how to reshape the healthcare system.” Healthcare informatics research 19.2 (2013): 79-85

[55] Akbari, Ali, et al. ”A flatter loss for bias mitigation in cross-dataset facial age estimation.” 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021.

[56] Hammond, M. Elizabeth H., et al. ”Bias in medicine: lessons learned and mitigation strategies.” Basic to Translational Science 6.1 (2021): 78-85.

[57] Mahabadi, Rabeeh Karimi, Yonatan Belinkov, and James Henderson. ”Endto-end bias mitigation by modelling biases in corpora.” arXiv preprint arXiv:1909.06321 (2019).

[58] Bender, Emily M., and Batya Friedman. ”Data statements for natural language processing: Toward mitigating system bias and enabling better science.” Transactions of the Association for Computational Linguistics 6 (2018): 587-604.

[59] Wang, Tianlu, et al. ”Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations.” Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[60] Maudslay, Rowan Hall, et al. ”It’s all in the name: Mitigating gender bias with name-based counterfactual data substitution.” arXiv preprint arXiv:1909.00871 (2019).

[61] Hort, Max, et al. ”Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods.” Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. 2021.

[62] Tae KH, Roh Y, Oh YH, Kim H, Whang SE. Data cleaning for accurate, fair, and robust models: A big data-AI integration approach. In Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning 2019 Jun 30 (pp. 1-4).

[63] Zhang L, Wu Y, Wu X. Causal modelling-based discrimination discovery and removal: criteria, bounds, and algorithms. IEEE Transactions on Knowledge and Data Engineering. 2018 Sep 30;31(11):2035-50.

[64] Hajian S, Domingo-Ferrer J. A methodology for direct and indirect discrimination prevention in data mining. IEEE Transactions on Knowledge and data engineering. 2012 Apr 3;25(7):1445-59.

[65] Elbassuoni S, Amer-Yahia S, Ghizzawi A, El Atie C. Exploring fairness of ranking in online job marketplaces. In22nd International Conference on Extending Database Technology (EDBT) 2019 Mar 26.

[66] Kamishima T, Akaho S, Asoh H, Sakuma J. Model-based and actual independence for fairness-aware classification. Data mining and knowledge discovery. 2018 Jan;32:258-86.

[67] P´erez-Suay A, Laparra V, Mateo-Garc´ıa G, Mun˜oz-Mar´ı J, G´omez-Chova L, Camps-Valls G. Fair kernel learning. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 2017 Dec 30 (pp. 339-355). Cham: Springer International Publishing.

[68] Kamishima T, Akaho S, Sakuma J. Fairness-aware learning through regularization approach. In2011 IEEE 11th International Conference on Data Mining Workshops 2011 Dec 11 (pp. 643-650). IEEE.

[69] Mancuhan K, Clifton C. Discriminatory decision policy aware classification. In2012 IEEE 12th International Conference on Data Mining Workshops 2012 Dec 10 (pp. 386-393). IEEE.

[70] Luong BT, Ruggieri S, Turini F. k-NN as an implementation of situation testing for discrimination discovery and prevention. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining 2011 Aug 21 (pp. 502-510).

[71] Kamishima T, Akaho S, Asoh H, Sakuma J. The independence of fairness-aware classifiers. In2013 IEEE 13th International Conference on Data Mining Workshops 2013 Dec 7 (pp. 849-858). IEEE.

[72] Zliobaite I, Kamiran F, Calders T. Handling conditional discrimination. In2011ˇ IEEE 11th international conference on data mining 2011 Dec 11 (pp. 992-1001). IEEE.

[73] Kamiran F, Karim A, Zhang X. Decision theory for discrimination-aware classification. In2012 IEEE 12th international conference on data mining 2012 Dec 10 (pp. 924-929). IEEE.

[74] Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: A call for open science. Patterns. 2021 Oct 8;2(10):100347.

[75] Prakhar, Jyoti, and Md Tanwir Uddin Haider. "Automated Detection of Biases within the Healthcare System Using Clustering and Logistic Regression." 2023 15th International Conference on Computer and Automation Engineering (ICCAE). IEEE, 2023.

[76] Rayavarapu, S. M. ., Prashanthi, T. S. ., Kumar, G. S. ., Lavanya, Y. L. ., & Rao, G. S. . (2023). A Generative Adversarial Network Based Approach for Synthesis of Deep Fake Electrocardiograms . International Journal on Recent and Innovation Trends in Computing and Communication, 11(3), 223–227. https://doi.org/10.17762/ijritcc.v11i3.6340

[77] Mr. Ather Parvez Abdul Khalil. (2012). Healthcare System through Wireless Body Area Networks (WBAN) using Telosb Motes. International Journal of New Practices in Management and Engineering, 1(02), 01 - 07. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/4