International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Automated Bias Detection within the Cardiovascular Disease Dataset using MapReduce Framework with Balance Measure

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

Submitted: 25/01/2024 **Revised**: 03/03/2024 **Accepted**: 11/03/2024

Abstract: Today, many fields rely on decision support systems, including health care for making appropriate decisions based on datasets. The decision support system, particularly in cardiovascular disease, is entirely dependent on the big data set, so if it is biased, it's difficult to decide whether the person has a cardiovascular disease. Bias detection in cardiovascular disease datasets has become a complex task because of the direct processing of large data sets. Another major drawback is that biases are detected on the set of attributes rather than protected attributes within the cardiovascular disease dataset which in turn increases computational cost as we know biases lie within protected attributes. Thus, it is a major challenge to identify the protected attribute from the set of attributes. Further, In the past bias identification was done manually using a statistical technique, which produced unreliable results i.e. minimum bias value related to cardiovascular disease. Considering all these challenges, we introduce a pioneering framework designed for automated bias detection within extensive cardiovascular disease datasets. Within our proposed methodology, we identified the protected attribute, namely gender, utilizing the capabilities of the MapReduce framework. Further, the balance measure approach has been used on the protected attribute of the cardiovascular disease dataset to detect the biases. The comparative results reveal that the detection of biases on protected attributes outperforms the existing works in terms of bias value, accuracy, precision, and F1 score which are 28%, 72%, 73%, and 81% respectively. These metrics collectively indicate the superior performance of the proposed methodology.

Keywords: *Bias, Big dataset, Balance measure approach, MapReduce Framework*

1. Introduction

In the present era, several industries, like healthcare and social media, heavily rely on big data for decision-making. Despite the extensive use of big data, key difficulties associated with big data include biased data, storage, security, and analytics [1]. Particularly in datasets related to cardiovascular disease, the presence of biases can contribute to the deterioration of data quality. Bias is a cognitive notion or assumption that hinders an individual's ability to make informed judgments based on information and study [2]. Nowadays, cardiovascular diseases (CVD) are a prominent source of morbidity and death across the world. Addressing this health challenge necessitates precise and unbiased data analysis approaches for analyzing data to facilitate rational decisions for healthcare systems. The healthcare system utilizes the big dataset for learning about disease patterns, predictions, and treatment outcomes. The fundamental problem, however, is assuring the quality and representatives of these datasets, especially when dealing with the numerous demographic and clinical factors associated with cardiovascular health. Biased cardiovascular disease (CVD) datasets can contaminate data integrity and diminish the decision-making capacity of systems, posing a societal risk. Therefore, detecting biases in models is crucial for enhancing performance [3]. These biases often stem from disparities or skewness in data, particularly associated with protected attributes such as gender, age, race, religion, and others, which are considered sensitive and require protection from discrimination.

Moreover, within the cardiovascular disease (CVD) dataset, gender is designated as the protected attribute, which can consequently result in biases. In the CVD dataset, biases related to gender can arise if the dataset is not representative of both genders or if there are inherent disparities in how certain conditions are diagnosed, treated, or reported for different genders. Therefore, if a dataset on cardiovascular disease is biased concerning gender, it may not accurately reflect the symptoms, prevalence, or outcomes of the disease for the respective genders. Considering these limitations, there is an urgent requirement to identify biases within the CVD dataset. Moreover, detecting biases also poses significant challenges and encompasses various complexities. i) The CVD depends highly on big datasets for decision-making which leads to an increase in time complexity due to the processing of big CVD datasets directly. ii) The evaluation of biases by different approaches has higher computational costs since it finds biases in the set of attributes rather than the protected attributes which is the major cause of the bias. iii) Detection of bias is performed

[1]*Dept. of CSE, National Institute of Technology Patna, Bihar – 800005, India*
Email: jyotiprakhar36@gmail.com, 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: jyotiprakhar36@gmail.com*

without knowing that a CVD dataset is in order or disorder state (the majority of one class is more than the other in the given attribute), which unnecessarily increases the computational cost. As we know biases generally exist in disordered conditions. Therefore, a methodology that can effectively recognize the disorder in the dataset is required. iv) In the past, biases were identified manually using statistical methods, resulting in inaccuracies and undermining the decision-making capabilities of cardiovascular disease analysis. Therefore, considering all these challenges, we have developed a robust framework for automated bias detection in cardiovascular disease datasets for this we have employed the MapReduce framework alongside a balance measure approach. By utilizing the parallel processing capabilities of the MapReduce framework, we aim to efficiently analyze big CVD datasets, to detect the bias that exists in the dataset. Moreover, the application of the disorder test involves utilizing the Shannon formula to ascertain the presence or absence of disorder within the system. Additionally, we applied a balance measure approach using the balance formula. This approach detects bias values in the cardiovascular disease dataset, specifically on protected attributes. The proposed methodology effectively overcomes important issues associated with bias detection in cardiovascular disease datasets and outlines recommendations for future research endeavours.

Motivation

In today's world, many applications rely largely on data-driven decision-making systems, which is especially obvious in the field of cardiovascular disease management. Ensuring equal access to high-quality healthcare remains a top priority, regardless of individual characteristics like race or gender. Unfortunately, hidden biases in cardiovascular disease diagnosis and treatment procedures can cause discrepancies in care quality, incorrect diagnoses, and diagnostic delays for patients. Furthermore, these biases may increase stress levels, aggravating existing medical issues. Researchers specializing in decision-making, data mining, and machine learning have launched initiatives to overcome these biases from several angles [4]. However, it is noteworthy that the bulk of attempts fall short of properly resolving the basic factors driving biased systems. Seeing this truth helps us to focus our efforts against biases in one of the major health-related issues which is cardiovascular diseases, and identify shortcomings in healthcare systems. Furthermore, addressing biases in cardiovascular disease datasets is critical for ensuring that the insights obtained from these datasets are relevant to various patient groups, hence enhancing customized and equitable treatment.

The sections of this research article are arranged as follows: Section II presents an overview of relevant literature, highlighting common issues and research questions. Section III describes the contributions contributed to this scientific endeavour. Section IV describes the mechanism for automatic bias identification. Finally, Section V summarises the study's findings and recommends future research possibilities.

2. Related Work

In this paper, to understand the recent works related to bias detection in health-related systems, we have investigated some of the papers which are as follows:

Kruse et al. [5], carried out a thorough literature analysis to investigate the difficulties and possibilities connected with big data in the healthcare sector. They emphasized the large volume of healthcare data created each year and the significance of categorizing and organizing this data to guarantee universal accessibility and transparency across healthcare facilities. However, the study's disadvantage is its dependence on a small number of articles, which may introduce biases and impair the clarity of the conclusions. According to the author of the research [6], [7], [8], one of the challenges in the healthcare sector is that their data is more unstructured than data from other fields. Furthermore, if the data is biased, the healthcare decision support system will be unable to reach an appropriate conclusion, which could impede performance and endanger society. Norori et al. [9], revealed that, while bias is ubiquitous in the medical area, measuring and identifying it can be difficult. The explosion of varied data sources that are constantly exchanged, acquired, and incorporated into artificial intelligence (AI) systems is shaping the ever-changing healthcare delivery environment. Furthermore, AI is positioned to provide data-driven techniques to strengthen clinical decision-making processes and promote public health efforts, gradually improving societal well-being. To stimulate improved cooperation between the medical and AI disciplines, as well as to provide a forum for varied viewpoints on the integration of AI in medicine, open scientific concepts must be incorporated into AI system design and assessment frameworks.

Zhao et al. [10], presented a novel bias evaluation and detection approach known as LOGAN (Local Group Bias Detection Algorithm), which is based on clustering. LOGAN uses a clustering technique to organize instances based on their properties, intending to optimize a bias measure (such as performance disparities across groups) inside each cluster. However, one disadvantage of this suggested framework is the potential for different cluster sizes. Furthermore, it detects bias using machine learning algorithms and performance measurements on

individual clusters rather than directly on protected properties, which may result in additional processing time. Lee et al. [11], studied the critical need to improve patient outcomes and healthcare quality, emphasizing the growing importance of data accessibility and analytical skills as the big data age in healthcare emerges. They stressed the need to increase the quality of data in electronic health records. However, one significant drawback of their analysis is the assumption that clinical practice easily integrates big data analytics, as this integration necessarily requires clear therapeutic advantages, as mentioned in earlier research [12], potentially resulting in increased time complexity. Zliobaite [13], surveyed to classify and examine several discrimination measuring methods that are used to analyze data bias and evaluate the effectiveness of discrimination-aware prediction algorithms. The author underlined how important it is to assess prediction models' fairness methodically and objectively. The study's main finding emphasizes that most previous research has been focused on binary classification problems using binary-protected features. Still, a significant shortcoming of this study is that it only considers statistical methods to quantify discrimination in data. Jena et al. [14], conducted a study to identify several sources of big data influence across many big data applications. The survey provided an assessment of the regression-based optimization technique for MapReduce applications, along with a blueprint for the scalable design and implementation of a clustering algorithm inside the MapReduce paradigm. The overall goal of future research endeavours mentioned in this paper is to improve Hadoop framework performance by reducing execution time.

Bhosale et al. [15], investigate the notion of big data, as well as the technological issues that must be addressed to handle data efficiently and quickly. However, one significant disadvantage noted is the existence of technological challenges that are economically prohibitive to handle within the boundaries of a particular application area. Bhathal et al. [16], examined many vulnerabilities in the Hadoop architecture and provided various solutions to mitigate or remove them. The experimental setting included performing common assaults to acquire insight into the idea and execution of preventative measures against such attacks. The findings highlight the impact of assaults on system performance. However, one key constraint is the need to secure data with defense-in-depth security techniques. Zhao et al. [17], provide a parallel k-means clustering technique based on MapReduce, which is well-known for its ease of use and effectiveness in parallel programming. The experimental results show that the suggested technique may easily scale and analyze big datasets on conventional hardware. However, one major disadvantage is the need to build dataset-oriented parallel clustering algorithms to improve performance even further. Desai et al. [18], present research that investigates the gender bias in cardiovascular disease, with an emphasis on coronary artery disease (CAD) prevention, detection, and therapy. It emphasizes gender inequities in research representation, diagnostic tools, and treatment procedures, highlighting the importance of gender-sensitive approaches. The study argues for a move toward inclusive research methodologies and revised clinical recommendations to address gender bias in CAD, and it serves as a valuable resource for healthcare professionals and policymakers working to achieve gender parity in cardiovascular treatment. Kim MD et al. [19], investigated the impact of gender bias as a mechanistic factor of cardiovascular disease outcomes. It explores how these biases lead to differences in illness appearance, progression, and treatment outcomes. The study emphasizes the need to identify and correct these biases to provide more accurate risk assessment and individualized cardiovascular treatment. It is a helpful resource for healthcare practitioners and academics working to improve the accuracy and equity of cardiovascular disease management.

Discussion

After carefully examining the literature listed in Section 2, we have identified critical areas necessitating further research to address existing gaps and enhance the overall performance of bias detection systems. A notable drawback in the current state of bias detection lies in the dependence on manual, statistical techniques, as discussed in [13]. This manual method is time-consuming and increases the possibility of errors. In addition, the research [11] identifies a noteworthy drawback that results from working directly with big datasets of health-related systems, which adds to increased temporal complexity. The inherent challenges of managing and processing extensive datasets further compound this drawback. Another notable limitation, as indicated by [10], is the absence of methodologies directly categorizing the protected attribute—a key source of biases in the dataset. This underscores the need for a systematic categorization of the specific protected attribute to effectively address biases within the dataset. Additionally, there is a pressing need to develop methods to quantify disorder within the dataset based on the protected attribute. According to several studies, this proactive method lowers computing costs and speeds the procedure compared to direct bias detection. For this reason, adding disorder measurement is a crucial step in creating a framework for bias detection that is both more effective and efficient.

2.1. Challenges

In Section 2, we provide the most recent findings to spark interest among researchers in the creation of biases inside health-related systems. Through highlighting shortcomings in the existing healthcare system, we want to bring attention to areas that require reform to move towards a more equitable state. By tackling these issues and offering solutions, we want to improve model performance and eventually contribute to a more equal healthcare environment.

- We've highlighted that relying solely on statistical techniques for bias detection in the CVD dataset yields unreliable results. Additionally, the manual approach lacks automation, potentially leading to errors and time-intensive processes.
- We have observed that the processing of large CVD datasets directly for the detection of biases significantly increases the time complexity, consequently degrading the overall performance of the system.
- We systematically surveyed that detection of biases is done on the set of attributes rather than protected attributes which is the main cause of the biases in the system. The detection of biases on the set of attributes will increase both the computational cost and time.
- We also identified that the detection of biases was performed on the CVD dataset without knowing that a dataset is in an order or disorder state, which unnecessarily increases the overall computational cost of the system. As we know biases generally exist in disordered conditions.

2.2. Research Problem

After carefully considering the substantial challenges noted in section 2.1, the research problem was developed i.e., automated detection of biases on protected attributes of cardiovascular disease dataset leverages the MapReduce framework by incorporating a balance measure approach.

3. Contribution

In this part, we will discuss major contributions that are essential for addressing the research problem at hand which are as follows:

- We have presented a methodology for automatically detecting biases in cardiovascular disease datasets. Our methodology takes the dataset as input and categorizes the protected attributes, which are frequently the source of biases. In addition to this, we use a balance measure approach to discover biases within the protected attribute. This will minimize the possible errors and reduce the processing time.
- We used K-Means clustering to cluster the big CVD dataset, significantly reducing the time spent processing the big data and increasing the overall performance.
- We've devised a methodology employing the MapReduce framework, which efficiently categorizes the protected attribute within the CVD dataset, thus optimizing computational time.
- We utilized Shannon entropy to quantify disorder within the CVD dataset. If disorder is detected, we employ the balance measure approach for bias detection, effectively reducing computational costs and expediting bias detection.

4. Methodology

In this section, we delve into the intricate details of a novel framework for automated bias detection in a cardiovascular disease dataset. To enhance the comprehensibility of the proposed methodology, we have developed a flowchart outlining its steps. Figure 1, below shows the flowchart which serves as a visual aid to elucidate the process and facilitate an understanding of how the bias can be detected within the CVD dataset.

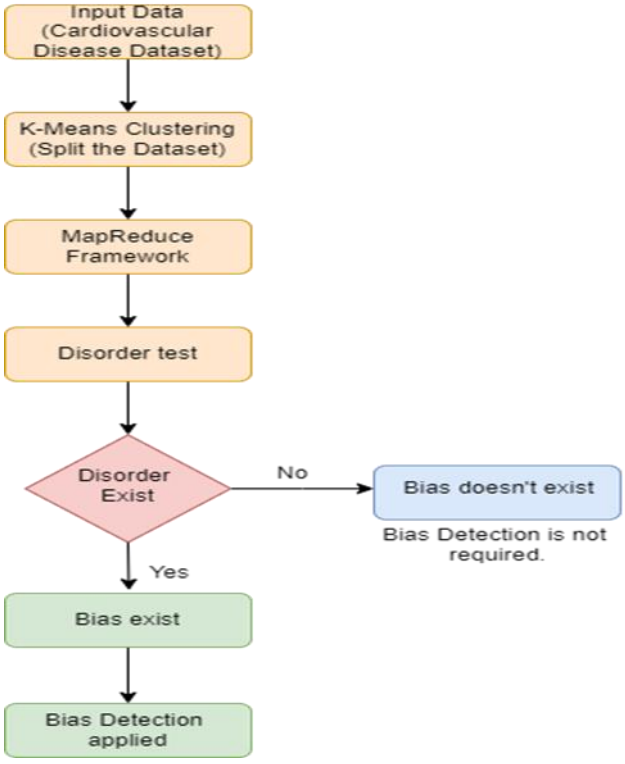

Fig. 1. Flowchart of Proposed Methodology

Furthermore, to implement the flowchart we have developed the proposed framework which is shown in Figure 2. This framework leverages the MapReduce framework [20] alongside a balance-measure approach. In the proposed framework, we have four modules as Clustering module, the MapReduce framework module, the disorder test module and the bias detection module.

The cardiovascular disease dataset has been given as input to the proposed framework, which is the big dataset, and further, the dataset is split into the form of clusters in the clustering module with the help of a clustering algorithm (K-Means clustering). After that, the MapReduce strategy is applied to the cluster data sets within the MapReduce Framework module to categorize the protected attribute and using the third module which is the disorder test module in which the disorder is tested on the categorized protected attribute by using the Shannon entropy and if the disorder is present then bias is detected by the bias detection module in which bias is detected by applying balance measure approach. The detailed description of this framework is outlined in the subsequent section.

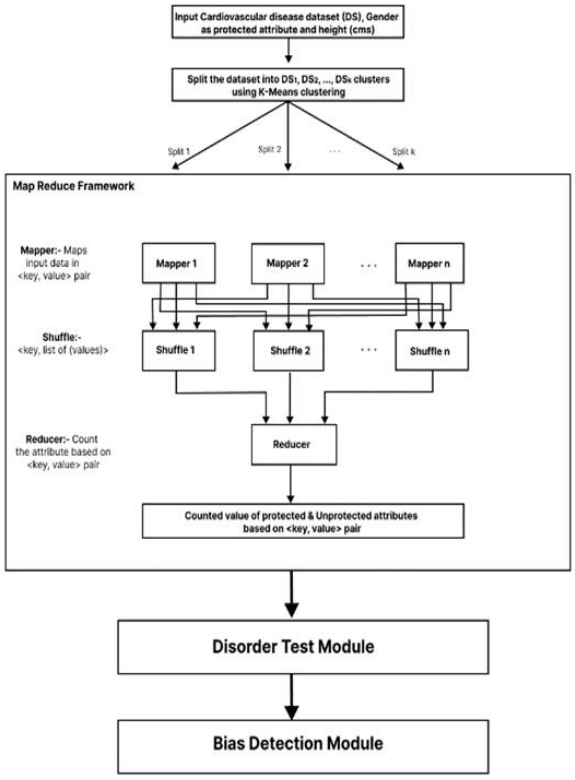

Fig. 2. The Proposed Methodology

4.1. Dataset

In this paper, for bias detection, the cardiovascular disease dataset (https://www.kaggle.com/datasets/sulianova/cardiovascular-diseas-edataset?resource=download) is used which is the big dataset. Figure 3 presents a snapshot of the cardiovascular disease dataset. In this dataset, there are 70000 rows and 13 columns such as id, age, gender, height, weight, ap_hi (Systolic blood pressure), ap_lo (Diastolic blood pressure), cholesterol, gluc (Glucose), smoke (Smoking), alco (Alcohol intake), active (Physical Activity), and cardio (Presence or Absence of Cardiovascular disease). Here, the protected attribute under consideration is gender, given the known differentiation in cardiovascular disease between males and females. For the clustering process, we've utilized two attributes from the cardiovascular disease dataset: gender (1 for female and 2 for male), serving as the protected attribute, and height (in centimetres), functioning as the unprotected attribute. Here we used two attributes only for clustering because clustering on a minimal attribute helps to identify the key factors easily that are driving the clustering results and gain insight data. Furthermore, the clustering algorithm is utilized on the input dataset, effectively partitioning the large dataset into smaller clusters.

data

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
69995	99993	19240	2	168	76.0	120	80	1	1	1	0	1	0
69996	99995	22601	1	158	126.0	140	90	2	2	0	0	1	1
69997	99996	19066	2	183	105.0	180	90	3	1	0	1	0	1
69998	99998	22431	1	163	72.0	135	80	1	2	0	0	0	1
69999	99999	20540	1	170	72.0	120	80	2	1	0	0	1	0

70000 rows × 13 columns

Fig. 3. The Snapshot of Cardiovascular Disease Dataset

4.2. Clustering

We performed clustering to gain insights into a population within a large CVD dataset. Employing the K-Means clustering method on the input dataset was imperative due to its vast size. Large datasets often present challenges such as feature complexity and insufficient diversity, potentially resulting in hidden biases. Detecting these biases becomes increasingly cumbersome as data volume escalates. This study's input dataset comprises 70,000 rows and 2 columns (gender, height). We applied K-Means clustering to partition the data into 10 clusters of varying sizes. Figure 4 illustrates the distribution of values within each cluster, with K=10 representing the number of clusters. Figure 5 showcases a snapshot of the output, displaying instances of gender and height within a single cluster.

```
[20] data_norm['clusters'].value_counts()
0s

0    7161
6    7143
5    7113
3    7048
2    7047
9    6990
7    6933
4    6892
1    6865
8    6808
Name: clusters, dtype: int64
```

Fig. 4. The value of Datapoints within each Cluster

For experimentation, we utilized an Intel(R) Core (TM) i7-4790 CPU @ 3.60GHz processor with 8.00 GB RAM, coupled with Hadoop version 3.2.3 and Java 1.8.0. With 10 nodes configured similarly, we established 10 clusters, assigning each node to handle a distinct cluster concurrently. This parallel processing approach significantly reduced the execution time for each cluster by dividing the input dataset into 10 partitions. Adjusting the number of clusters remains flexible, dependent on resource availability and specific requirements. Following clustering, the clustered output seamlessly integrates into the MapReduce framework for subsequent processing.

Fig. 5. The various instances of gender and height

4.3. MapReduce Framework

The output from the clustering algorithm is then fed into the MapReduce framework module. In the MapReduce paradigm, the "reduce" step follows the "map" job, aiming to minimize processing power and cluster network overhead. Initially, each job is mapped before being reduced into equivalent tasks. The MapReduce framework architecture, as depicted in Figure 6, is applied to clustered datasets of cardiovascular disease. The input to the MapReduce framework comprises the clustered dataset, consisting of attributes such as gender (1 for female, 2 for male) and height (in cms). The mapper function transforms the input data into ⟨Key, value⟩ pairs. Here, the keys represent gender (1 for female, 2 for male), while the value is set to 1 for each occurrence. For instance, ⟨1, 1⟩ denotes a female occurrence, where key=1 represents females and value=1 indicates the frequency of female instances. Subsequently, the ⟨Key, value⟩ pairs are shuffled into ⟨Key, a list of (values)⟩ format before being sent to the reducers. For instance, ⟨1, (1, 1) ⟩ signifies that key=1 represents females, and the list of values= (1, 1) corresponds to the occurrences of females. Lastly, the reducer consolidates the input data into ⟨Key, Total values⟩ pairs, tallying the overall sum of keys. For example, ⟨1, 4⟩ indicates that key=1 represents females, and the total count of females across all clusters is 4.

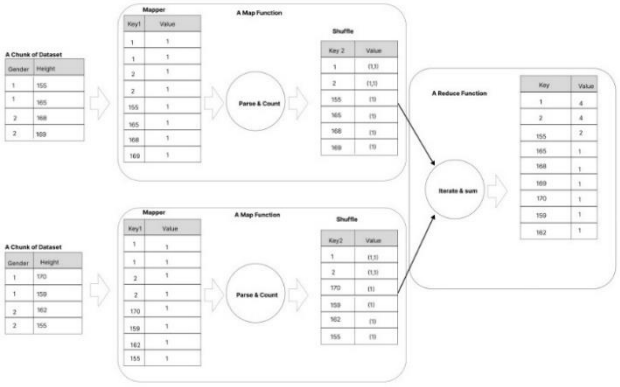

Fig. 6. The MapReduce framework

Figure 7 below illustrates the output of the MapReduce framework for the input clustered cardiovascular disease dataset. This output provides comprehensive counts for the attributes, specifically gender and height. In the dataset, the frequency of females (represented by 1) is 45,530, while the frequency of males (represented by 2) is 24,470. We focus solely on the protected attributes, namely gender (females and males), hence the overall counts for females and males are provided above. Additionally, the frequency gender, and height corresponding to each key are depicted using a bar chart in Figure 8.

1	45530
150	1051
151	613
152	1161
153	1059
154	1443
155	1781
156	2755
2	24470
157	1813
158	3313
159	1994
160	5022
161	1712
162	3256
163	2516
164	3396
165	5852
166	1979
167	2537
168	4398
169	2791

Fig. 7. The Output of MapReduce Framework

Subsequently, the output from the MapReduce framework is forwarded to the disorder test module for testing the disorder on the categorized protected attribute by using the Shannon entropy and if the disorder is present then identify the value of biases in the dataset.

4.4 Disorder Test Module

In this module, we test the disorder on the protected attribute of the dataset, and if the disorder is detected, then we proceed to apply the balance measure approach in the bias detection module to quantify the percentage of bias.

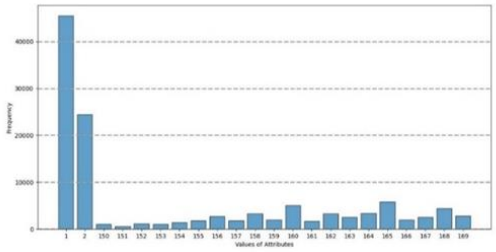

Fig .8 The Gender and Height Frequncies

For the disorder test, we incorporate a mathematical model involving Shannon entropy (H) to quantify the disorder within the dataset. The Shannon entropy, ranging from 0 to 1, tends towards 0 in cases of disorder. It evaluates the disorder using the number of instances in the dataset (n), the number of classes in the protected attribute (k), and the size of each class (c_i). In our study, the k represents the gender. The output of the MapReduce framework, focusing on the protected attributes (female and male gender), is passed to the Shannon entropy for the disorder test.

The test is done based on equation (1) which is Shannon Entropy(H):

$$H = -\sum_{i=1}^{k} \frac{c_i}{n} \log \frac{c_i}{n} \quad (1)$$

Where $n =$ instances in the dataset,

$k =$ number of classes,

c_i = size of each class.

$H = 0$, if there is only one class.

$H\ tends\ to\ 0$ when your data collection is severely uneven.

The disorder is calculated using the Shannon entropy where the value of k is 2 (for male and female), c_i for the male class is 24470 and c_i for the female class is 45530 respectively and the value for n is 70000. Hence, the calculated value H i.e. **Shannon entropy is 0.28**, which is shown in Table 1 and it lies within the threshold values 0 and 1. The value of H tends to 0 which means disorder exists in the dataset. Once the disorder is tested it finally goes to our last module i.e. bias detection module to measure the bias value in the CVD dataset.

Table 1. Disorder Measurement in Dataset

Disorder Test	Value
Shannon Entropy	0.28

4.5. Bias Detection Module

In this module, we assess the extent of bias on the protected attribute present in the dataset by applying the balance measure approach. Following the disorder assessment, the balance measure approach is employed to determine the percentage of bias. This balance measure is utilized particularly when one class within the protected attribute is disproportionately prevalent compared to others, leading to bias within the system. The bias detection is done using equation (2) which is a balance measure.

$$Balance = \frac{H}{\log k} = \frac{-\sum_{i=1}^{k} \frac{c_i}{n} \log \frac{c_i}{n}}{\log k} \quad (2)$$

Where H = Shannon Entropy,

k = number of classes.

Table 2. Detection of Bias Value for Gender

Balance Measure Approach	Value
Balance Formula	28%

The above Table 2, reveals that the dataset exhibits a bias of 28% towards females. This underscores the presence of biases within the dataset. Where, H is Shannon entropy which was calculated previously as 0.28 and k is the number of classes in protected attributes which is 2 (gender i.e., Male and Female). Hence, the bias value by using the **balance measure is 28%** which shows that the CVD dataset is biased. Further, we conducted a comparative analysis of our results with those presented in another research paper referenced in [21], as depicted in Table 3.

Table 3. Comparison of the Results

Paper	Techniques	Disorder Test	Detection of Biases	Bias Value	Acc.	Precision	F1-Score
Suri et al. [21]	Classification using Machine	Not Performed	On the set of attributes	20.20%	44.20%	20.80%	8.20%

Proposed Methodology	Learning Clustering, MapReduce Framework and balance measure approach	Performed	Only on Protected Attribute	28%	72%	73%	81%

The comparison table above highlights a notable distinction between our proposed methodology and the approach described in [21]. While [21] relies on machine learning classification applied to a set of attributes, resulting in increased time complexity and degraded performance, our method employs clustering followed by the MapReduce framework. This approach individually classifies the dataset based on specific protected attributes, thereby minimizing time complexity and enhancing performance. Moreover, our methodology incorporates a disorder test using Shannon entropy (H) within the dataset. This ensures efficiency by only proceeding to bias detection if the disorder is detected ($H \neq 0$), thus saving time compared to [21], where bias detection is performed directly without the disorder test. For assessing the imbalance factor in the dataset, we employ the balance measure approach to quantify the bias value. Our analysis reveals a bias value of 28%, significantly higher than that obtained in [21]. Furthermore, we evaluate multiple performance metrics, including accuracy, precision, and F1-score, yielding values of 72%, 73%, and 81%, respectively. These results underscore the superior performance of our proposed method compared to [21].

5. Conclusion and Future Work

In our work, we introduce a novel framework for the automated detection of bias in cardiovascular disease datasets, crucial for enhancing model performance. Leveraging the MapReduce framework and various balance measure approaches, our methodology directly classifies the protected attribute within the dataset. Shannon entropy is then utilized to quantify disorder in the dataset based on the protected attribute, yielding a calculated value of 0.28 in this instance. Subsequently, applying the balance formula using Shannon entropy reveals a bias measure of 28%, indicating the presence of gender bias in the dataset due to the unbalanced distribution of the protected attribute. These findings highlight the disparity within the system, ultimately impacting overall performance. Comparison with other research papers emphasizes the effectiveness of our methodology, as bias detection is conducted based on protected attributes, saving time by focusing solely on relevant factors. Additionally, we evaluate compelling performance metrics, with accuracy, precision, and F1-score achieving notable values of 72%, 73%, and 81%, respectively, further affirming the superiority of our approach. Moving forward, future endeavors may involve identifying the specific types of biases present in the system and implementing mitigation strategies to promote fairness. Furthermore, optimizing a fair prediction model remains a key area for research, aimed at fostering more accurate and equitable predictive models to enhance patient care within cardiovascular research.

Acknowledgments We extend our heartfelt gratitude to Dr. Md. Tanwir Uddin Haider for his unwavering dedication, continuous encouragement, and invaluable guidance throughout this research endeavor. His exceptional support has been instrumental in shaping the trajectory of this study and fostering our academic development. We deeply appreciate Dr. Haider's profound contributions, which have significantly enriched our work and propelled it to its fullest potential. His mentorship has left an enduring impact on both this research and our personal growth, for which we are sincerely grateful.

Author contributions

Jyoti Prakhar: Implementation, Proposed Novel Framework, Writing an original draft.

Dr. Md. Tanwir Uddin Haider: Identified Challenges and Conceptualization.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] A. Ghosh, Big data and its utility, Consulting Ahead 10 (2016) 52–69.

[2] Prakhar, Jyoti, and Md Tanwir Uddin Haider. "Bias Detection and Mitigation within Decision Support System: A Comprehensive Survey." International Journal of Intelligent Systems and Applications in Engineering 11.3 (2023): 219-237.

[3] 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.

[4] Agathe Balayn, Christoph Lofi, and Geert-Jan 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, 30(5):739–768, 2021.

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

[6] 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.

[7] 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.

[8] 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.

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

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

[11] Lee, Choong Ho, and Hyung-Jin Yoon. "Medical big data: promise and challenges." Kidney Research and clinical practice 36.1 (2017): 3.

[12] Rumsfeld, John S., Karen E. Joynt, and Thomas M. Maddox. "Big data analytics to improve cardiovascular care: promise and challenges". Nature Reviews Cardiology 13.6 (2016): 350-359.

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

[14] Jena, Bibhudutta, et al. "A survey work on optimization techniques utilizing map-reduce framework in Hadoop cluster." International Journal of Intelligent Systems and Applications 9.4 (2017): 61.

[15] Bhosale, Harshawardhan S., and Devendra P. Gadekar. "A review paper on big data and Hadoop." International Journal of Scientific and Research Publications 4.10 (2014): 1-7.

[16] Bhathal, Gurjit Singh, and Amardeep Singh. "Big data: Hadoop framework vulnerabilities, security issues and attacks." Array 1 (2019): 100002.

[17] Zhao, Weizhong, Huifang Ma, and Qing He. "Parallel k-means clustering based on MapReduce." Cloud Computing: First International Conference, CloudCom 2009, Beijing, China, December 1-4, 2009. Proceedings 1. Springer Berlin Heidelberg, 2009.

[18] Desai, Shailesh, Atul Munshi, and Devangi Munshi. "Gender bias in cardiovascular disease prevention, detection, and management, with specific reference to coronary artery disease." Journal of mid-life health 12.1 (2021): 8.

[19] Kim, Isabel, et al. "Sex and gender bias as a mechanistic determinant of cardiovascular disease outcomes." Canadian Journal of Cardiology 38.12 (2022): 1865-1880.

[20] Park, Dongchul, Jianguo Wang, and Yang-Suk Kee. "In-storage comput- ing for Hadoop MapReduce framework: Challenges and possibilities." IEEE Transactions on Computers (2016).

[21] Suri, Jasjit S., Mrinalini Bhagawati, Sudip Paul, Athanasios Protogeron, Petros P. Sfikakis, George D. Kitas, Narendra N. Khanna et al. "Under- standing the bias in machine learning systems for cardiovascular disease risk assessment: The first of its kind review." Computers in biology and medicine (2022): 105204.