Fairness Issues, Current Approaches, and Challenges in Machine Learning Models

Tonni Das Jui and Pablo Rivas

Computer Science, Baylor University, One Bear Place 97356, Waco,
76798, Texas, USA.

*Corresponding author(s). E-mail(s): Tonni_Jui1@Baylor.edu;
Contributing authors: Pablo_Rivas@Baylor.edu;

Abstract

With the increasing influence of machine learning algorithms in decision-making processes, concerns about fairness have gained significant attention. This area now offers significant literature that is complex and hard to penetrate for newcomers to the domain. Thus, a mapping study of articles exploring fairness issues is a valuable tool to provide a general introduction to this field. Our paper presents a systematic approach for exploring existing literature by aligning their discoveries with predetermined inquiries and a comprehensive overview of diverse bias dimensions, encompassing training data bias, model bias, conflicting fairness concepts, and the absence of prediction transparency, as observed across several influential articles. To establish connections between fairness issues and various issue mitigation approaches, we propose a taxonomy of machine learning fairness issues and map the diverse range of approaches scholars developed to address issues. We briefly explain the responsible critical factors behind these issues in a graphical view with a discussion and also highlight the limitations of each approach analyzed in the reviewed articles. Our study leads to a discussion regarding the potential future direction in ML and AI fairness.

Keywords: ethics, model fairness, bias reduction, fair prediction, AI, machine-learning

1 Introduction

Machine learning-based models have undoubtedly brought remarkable advancements in various fields, demonstrating their ability to make accurate predictions and automate decision-making processes. However, many real-world applications of machine learning (ML) models, such as determining admission to a university [1], screening job applicants [2–4], disbursing government subsidies [5, 6], identifying persons at high risk of disease [7], and so on, are prone to bias. The inherent biases and limitations in the training data and algorithms can lead to discriminatory outcomes and perpetuate societal biases. Discriminatory outcomes refer to situations where machine learning models produce predictions or decisions that systematically favor or disadvantage particular groups more than others [8, 9]. Societal biases are the preconceived notions, prejudices, or stereotypes in a society that can lead to unfair advantages or disadvantages for specific individuals or groups. ML and AI researchers have raised many questions about the source and the solution to the fairness issues in AI (FAI) and discussed various types of biases [10].

ML models can exhibit various unfairness issues, encompassing biases and discriminatory outcomes. Discussions often revolve around biases in training datasets, discriminatory behavior exhibited by predictive models, and the challenge of interpreting and explaining the outcomes produced by these models. Biases in the training dataset usually refer to the data representing disparities and discrimination against certain groups based on attributes such as race, gender, or socioeconomic status, which Ml models may inadvertently amplify. The press and literature gradually started to discuss these types of ML model bias in the early twenty-first century [11, 12]. Also, ML models can exhibit bias towards specific groups despite unbiased training data. Other than these issues, the prediction outcome's unexplainable and uninterpretable nature is another widespread fairness issue. Explainability and interpretability refer to the logical reasoning of outcomes with available alternative profiles. For example, suppose a person is denied credit from a bank. In that case, explanations provide feedback on where exactly his profile could be altered to get the credit, such as increasing monthly income, decreasing loan amount, or changing race. Some of these changes may not be possible, such as changing a person's race to get credit from a back, which makes the bank's credit assigning model unfair or biased towards a group of people [13]. In addition, researchers also report some other forms of bias, including inconsistent predictions and inherent biases within the data [14, 15].

As these bias types can compromise the integrity and reliability of decision-making procedures, impeding the advancement the ML model originally intended to enable, achieving fairness in ML predictions is essential [16]. Avoiding fairness concerns across diverse domains, including sign language analysis (e.g., [17, 18]), image object analysis (e.g., [19, 20]), non-linear data analysis (e.g., [21, 22]) and graph data analysis (e.g., [23]), could provoke doubts regarding the credibility and reliability of machine learning methodologies in respective fields. Advancement in ensuring fairness requires continuous research, development, and implementation of approaches to mitigate predictions' discrimination. Numerous researchers have discussed and proposed various approaches to this ongoing challenge in recent years, leading to rapid and dynamic growth in research within the field [24–28]. As a consequence of this growth, comprehending the existing issues and methodologies within the field can be time-consuming, highlighting the need for dedicated efforts to stay up-to-date with the latest advancements. It even requires much effort for people new to the field as a researcher. Literature review articles aid in this situation and provide comprehensive information so that researchers and practitioners can understand the proposed methodologies and their limitations with minimal effort. Also, it allows for examining different fairness definitions, evaluation metrics, and bias mitigation strategies employed in various domains. Moreover, a literature review helps identify gaps, challenges, and open research questions in the pursuit of fairness, enabling researchers to build upon existing work and propose novel approaches. Additionally, it aids in creating a shared knowledge base and promotes collaboration within the research community, ultimately contributing to developing more robust, transparent, and equitable machine learning models.

Although there are many literature review articles on fairness-ensuring approaches, some limitations persist in these works. Firstly, many studies need more discussion regarding the article exploring and collecting process [29–32]. Secondly, recent methodologies presented in these articles may need to be updated as researchers continue advancing the field [29]. In this regard, it is common for some approaches to lose relevance and for new approaches to gain significant impact, shaping the direction of research in machine learning and AI. Therefore, staying updated with the latest advancements is essential to ensure continued progress and relevance. Thirdly, although the usual goal of fairness-related articles is to generalize fairness definitions from various perspectives and develop an approach where this defined fairness is ensured, some literature review articles highlight various fairness definitions more than developed fair approaches [29]. However, understanding the procedures to ensure fairness is as crucial as comprehending the various fairness-related terminologies. Lastly, there is a need for a more standardized evaluation and classification of fairness methodologies from the perspective of their addressed fairness issues. Most reviews classify fairness-ensuring methodologies based on when the researchers are incorporating a bias mitigation strategy (Prior to the model implementation, after the model implementation, or during the model implementation). We need to connect these fairness-ensuring methodologies with the specific issue types. Emerging academics often require more direction for understanding a classification of methodologies from the perspective of specific fairness issues they solve. Researchers often adhere to conventional methodologies when addressing specific challenges in their field. For example, scholars usually explore debiasing techniques for removing inherent data bias and generate counterfactual examples to explain model prediction.

To solve these issues, we offer a comprehensive mapping analysis of some recent fairness concerns and academics' proposed strategies. A comprehensive mapping study can present a clear idea of how to explore this field of research, which is especially helpful for aspiring scholars. Besides, a mapping study is a valuable tool for identifying and retrieving recent articles, facilitating the collection of related studies in subsequent years, even if the discussed articles become outdated. In this regard, the impact of a mapping study can endure longer than that of conventional review articles. Fig. 1 graphically represents our motivation to follow a systematic mapping study. Our paper also generalizes fairness-related terminologies with appropriate examples and the adopted approaches for ensuring them. Finally, we also present a taxonomy of fairness-ensuring methodologies from the perspective of fairness issues they solve. This discussion can be a good source for identifying new trends in proposed solutions, their

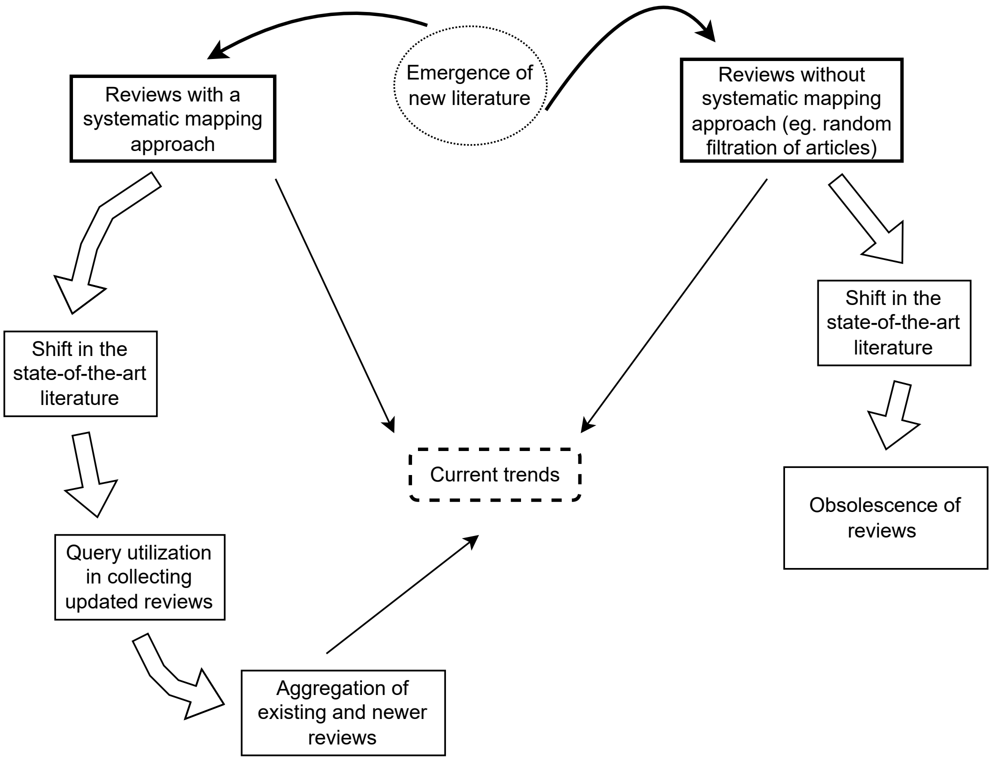
Emergence of new literature
Reviews with a systematic mapping approach
Reviews without systematic mapping approach (eg. random filtration of articles)
Shift in the state-of-the-art literature
Shift in the state-of-the-art literature
Current trends
Obsolescence of reviews
Query utilization in collecting updated reviews
Aggregation of existing and newer reviews

Fig. 1 Motivation to follow systematic mapping approach visualized in the diagram's sequential phases (contained within square boxes) from top to bottom. Comprehensive review articles assist us in learning current trends. Over time, the reviews with a systematic mapping approach and reviews without a systematic mapping approach are affected by the emergence of new literature. The article collection process aids in consolidating emerging information with existing reviews, enabling continued comprehension of contemporary trends, even over an extended period.

limitations, and potential future directions in state-of-the-art articles. We summarize our contributions as follows,

- We offer a systematic mapping study of 94 articles that address fairness concerns, bias mitigation strategies, fairness terminology, and metric definitions across distinct research groups. Our mapping study method presents an adaptable search query for multiple databases that explains the article filtering process and an overview of how the number of articles increased/decreased over time and which countries are mostly involved in these 94 articles. This description of the filtering process facilitates the credibility of the work. Also, new researchers can follow or tune the query to review more updated papers for an extended period.
- We classify the fairness issues first and then further classify the fairness-ensuring methodologies adopted to solve each type of fairness issue. We also represent the graphical taxonomy of fairness issues in Fig. 6, the taxonomy of methodologies in Fig. 7, and a taxonomy representing specific methodologies targeted to solve specific

issues and their limitations in Fig. 10 for researchers to understand this area's current trends easily.
- We describe each type of fairness issue, summarize the approaches described in the filtered articles and discuss their limitations. We also provide a detailed definition of the fairness terminologies explored in the filtered articles for applying in the bias mitigation approaches, explain them with related examples, and provide the metric definition of the fairness terminology (if available).
- We provide ideas for future contributions that researchers have yet to explore.
- We summarize the publicly available datasets that other researchers have explored in the filtered articles and open-source tools that other researchers have proposed for mitigating bias or identifying bias in a dataset.

We organize the rest of the paper as follows: Section 2 presents the background material necessary to follow our discussion. Then, section 3 discusses our methods of this mapping study along with the research questions we are attempting to answer in this article and the developed query. Section 4 represents the findings from our mapping study in the form of answers to current research trends and most engaged countries and individuals, the first two research questions mentioned in section 3. The rest of the research question answers are regarding the analysis of the filtered papers. The research problems discussed in the filtered papers, the adopted methodologies to solve them, and the limitations or challenges of these methodologies are in section 5, 6, 7 accordingly. Next, the following two sections, 8 and 9, represent the answers to the last two questions regarding future direction and publicly available datasets, tools, and source code. We also discuss threats to the validity of our work in section 10. Finally, we conclude in section 11 with a general summary of our contributions.

2 Background

Several notable literature review articles have examined the landscape of fairness-ensuring methodologies. Some articles generalized explanations of bias types and their sources from various perspectives. For example, Reuben Binns discussed two types of unfairness issues from the perspective of discrimination against groups: algorithmic discrimination against protected feature groups and lack of individual fairness in the algorithm as the bias [29]. Algorithmic discrimination against protected feature groups refers to the situation where machine learning algorithms result in unfair treatment or unfavorable outcomes for certain groups of individuals based on their protected attributes, such as race, gender, or age. Lack of individual fairness in algorithms refers to the situation where the algorithm treats similar individuals differently, leading to unfair outcomes. This fairness issue is problematic because it can result in unjust outcomes for individuals who are similar in relevant respects but are treated differently by the algorithm. Both types of unfairness can arise when the algorithm uses inappropriate features or biased training data to make decisions. Unlike Reuben Binns, Mehrabi et al. discussed biases from the perspective of the source of the bias. They explained three types of ML model biases: training data bias, algorithm bias, and user-generated data bias [31]. Here, training data bias refers to biases in the data used to train machine learning models, resulting in models that reflect and reinforce the biases present in the data. Next, Algorithm bias refers to the bias that may be introduced into machine learning models by the algorithms used to train them. This bias can result from the selection of a particular algorithm or from how we implement the algorithm. Finally, User-generated data bias refers to the bias when we train the model with user-generated data that may reflect the biases and preferences of the users who generated it rather than being representative of the population as a whole.

These review articles emphasize discussing the adopted fairness-ensuring methodologies and often classify these methodologies. Generally, they classify these methodologies into pre-processing, in-processing, and post-processing [30, 31]. Simon Caton organized a taxonomy with these classes and subdivided them further to lead a conversation on current methodologies [30]. Firstly, Pre-processing methods involve manipulating the training data before feeding it into the machine learning algorithm. This process generally involves data cleaning, feature selection, feature scaling, or sampling methods to ensure the data is balanced and representative of the population. Examples of pre-processing methods include data augmentation and demographic parity-ensuring methods. Data augmentation indicates data modification to balance underrepresented classes, and demographic parity-ensuring strategies indicate equalizing the proportion of positive outcomes across different protected groups. Secondly, in-processing methods modify the machine learning algorithm during the training process to ensure fairness. These methods involve modifying the objective function or adding constraints to the optimization problem to ensure a fair outcome from the model. Examples of in-processing methods include adversarial training, where a separate model predicts the protected attribute and the original model ensures that it does not use this information to make predictions, and equalized odds, where the algorithm is optimized to ensure that the true-positive rates and false-positive rates are equal across different protected groups. Lastly, the post-processing methods involve modifying the output of the machine learning algorithm to ensure fairness. These methods involve adding a fairness constraint to the output, adjusting the decision threshold, or applying a re-weighting scheme to the predictions to ensure they are fair. Examples of post-processing methods include calibration and reject option classification. Calibration in machine learning refers to adjusting a model's output to match the true probability of an event occurring better. A reject option allows the model to abstain from predicting uncertain inputs rather than making a potentially inaccurate prediction. Overall, these three categories and taxonomies of methods provide a range of options for researchers and practitioners to address bias and discrimination in machine learning models.

Along with leading a discussion regarding issues and methodologies, these articles represent fairness-related terminologies and metrics. In this regard, Chen et al. categorized fairness definitions into two groups: individual fairness and group fairness [32]. Individual fairness entails treating similar individuals equally, regardless of their group membership. In contrast, group fairness aims to ensure that the model treats different groups of individuals fairly, regardless of their protected attributes such as gender, age, or race. This type of fairness ensures that the algorithm does not discriminate against any specific group. Besides them, Mehrabi et al. proposed a more granular concept of fairness, called subgroup fairness, that focuses on ensuring fairness for relevant subgroups of individuals based on protected attributes and other relevant factors [31]. This fairness involves identifying subgroups of individuals with particular characteristics and ensuring that the algorithm treats them fairly. Another vital aspect of fairness-ensuring methodologies is measuring the degree of fairness a model achieves. To this end, scholars have proposed various fairness metrics that quantify different aspects of fairness. One review article comprehensively discussed these metrics and categorized them into different types [30]. The first category is abstract fairness metrics, based on mathematical properties such as independence, separation, and sufficiency. The second category is group fairness metrics, which measure how well the algorithm performs for different groups of individuals based on their protected attributes. Finally, the fourth category is individual and counterfactual fairness metrics, which consider the hypothetical scenario of how the model would have behaved if specific protected attributes were different. These fairness definitions and metrics are crucial in evaluating the performance of fairness-ensuring methodologies and can guide the development of algorithms that achieve the desired level of fairness.

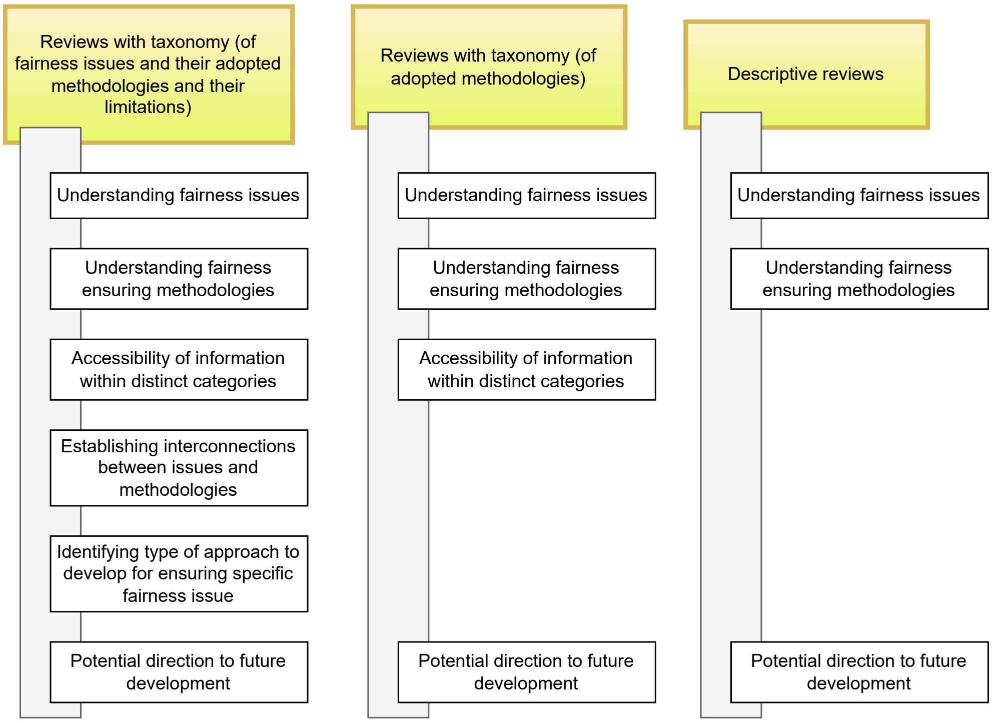
Reviews with taxonomy (of fairness issues and their adopted methodologies and their limitations)
Understanding fairness issues
Understanding fairness ensuring methodologies
Accessibility of information within distinct categories
Establishing interconnections between issues and methodologies
Identifying type of approach to develop for ensuring specific fairness issue
Potential direction to future development
Reviews with taxonomy (of adopted methodologies)
Understanding fairness issues
Understanding fairness ensuring methodologies
Accessibility of information within distinct categories
Potential direction to future development
Descriptive reviews
Understanding fairness issues
Understanding fairness ensuring methodologies
Potential direction to future development

Fig. 2 Aspects of various review structures.

Although the reviews offer valuable insights into different aspects of fairness, their limitations indicate the need for a more systematic and organized mapping study. For example, one significant limitation of these reviews is the need for more discussion on the interrelationships between the classification of fairness issues and the appropriate fairness-ensuring methodologies. While these reviews with methodology-based taxonomies/classifications bring synthesized insights into discussing the current fairness trends [29–32], it is difficult to link up these methodology classifications with the fairness issues that they solve and with the issues that they generate themselves. For academics, it is crucial to acknowledge that different types of fairness issues may require different types of methodologies for mitigation. Thus, a more detailed exploration of the links between the classification of fairness issues and the corresponding fairness-ensuring methodologies could facilitate the development of more effective and tailored solutions to address fairness issues in machine learning. Understanding the interconnection between issue groups and adopted method groups enables researchers to learn the appropriate method types they need to develop for a specific issue. Fig. 2 depicts a graphical representation of different aspects of various review structures that motivate our study. In this context, we propose taxonomy delineating pivotal factors that give rise to diverse classes of fairness concerns. Additionally, we categorize the methodologies employed to address each issue class and outline their respective limitations. By establishing these connections between fairness issue groups, corresponding resolution approaches, and their constraints, our taxonomy provides a comprehensive overview of prevailing trends within this domain.

3 Method of Mapping Study

We followed mapping techniques from other articles to analyze the major research trends in Ethical Machine Learning over the past two decades [33, 34]. Our mapping techniques involve identifying relevant publications by conducting a comprehensive search of four major databases, including ACM DL, IEEE Xplore, SpringerLink, and Science Direct, focusing on papers on the fairness concept. We selected these databases because they are widely renowned within the research community. To ensure a systematic approach, we followed the search and selection process recommended by B. Kitchenham [33, 34] and structured our research queries on key subject phrases and synonyms of those words for various indexing sites based on the process described by D. Das et al. [35]. Finally, we filtered the studies based on their relevance to our goal. We represent the filtered studies regarding year, countries, and authors. We elaborately describe these steps below.

3.1 Research question development and refinement:

This mapping study investigates how ethical AI and ML model researchers developed and utilized approaches to mitigate bias. We directed our efforts toward structuring and refining the research inquiries in alignment with Creswell et al.'s guidelines [36]. Drawing from Creswell et al. and Wayne et al., our approach involved formulating research questions comprising a primary inquiry accompanied by subsidiary inquiries [36, 37]. This methodology mirrors the approach adopted by previous researchers in constructing pivotal central and subsidiary questions pertinent to their objectives [35, 38, 39]. Our overarching objective revolves around delving into state-of-the-art research on fairness concerns, prompting an exploration of the involved researchers and their geographic affiliations within this research domain. This exploration encompasses their resolved challenges, methodologies employed, prospective research avenues, and experimental tools or datasets. Consequently, we have formalized the ensuing research questions (RQs):

i. What is the state-of-the-art research on fairness issues in AI?
ii. Which countries and individuals are the most engaged in this field of study?
iii. What problem have they solved?
iv. What method have they adopted to solve that problem?
v. What are the next challenges of the research?
vi. What is the future direction?
vii. Did they provide the source code and the dataset?

We maintained documentation while reading articles and determining the answers to the questions mentioned above for each article.

3.2 Query design:

Our query development process involves breaking down the study subject into a few key phrases. Then we attempted other combinations of synonymous words to those phrases. There were a total of three segments in our search query. First, we included the search phrase “Artificial Intelligence” in the initial segment of our query. We also included similar terms such as “AI”, “ML”, and “Machine Learning” in that portion. Next, we considered keywords, such as model, prediction, outcome, decision, algorithm, or learning for the second segment, as we wanted to explore the articles focusing on fairness ensuring only for ML models. In the third segment, we used concepts synonymous with ethical fairness or bias, such as fairness, fairness, ethics, ethical, bias, discrimination, and standards, to narrow our search results. Finally, for the last segment, we chose ‘mitigating bias’, ‘bias mitigation’, ‘removing bias’, ‘bias removal’, ‘fairness definition’, ‘explanation’, and ’interpretation’ keywords. Fig. 3 depicts the search query.

```
(ML OR Machine learning OR AI OR artificial intelligence)
AND (model OR prediction OR outcome OR decision OR
algorithm OR learning)
AND (fairness OR fair OR ethics OR ethical OR bias OR
discrimination OR standards)
```

Fig. 3 The designed query consists of three main parts for collecting articles from various databases with relevant keywords

3.3 Article collection, organization, filtering and mapping:

We selected research based on our search query, and our search query yielded a significant number of articles. Nevertheless, only some of these articles were within the

Table 1 Search Query Results for Indexer Sites

Database	Search Result	Filtered	Referenced	Total	List of Papers
IEEE Xplore	74	8	2	10	[25, 26, 40–47]
ScienceDirect	90	3	3	6	[48–53]
Springer Link	135	9	3	12	[54–66]
ACM DL	121	19	9	28	[67–94]
Other	NA	NA	38	38	[27, 28, 73, 95, 95–130]
Total	420	25	8	94	All of the above

scope of our research. During the initial screening, we filtered these papers. Following, we applied targeted screening approaches to filter out publications with insignificant impact on this subject, excessive length, published in languages other than English, and repeated or identical research. We also attempted to see if the full text of the article was available publicly and if the author's claim was well-referenced. Finally, we analyzed the results from our search query for multiple ranges. For example, we attempted to find our query terms in the paper's 'abstract', 'introduction', 'conclusion' and 'title' or 'anywhere in the text', and so on. Then, we looked through the related work section of the remaining review articles, adding significant research that the search query had missed. In case these additional papers belong to any of the four databases (IEEE Xplore, ACM DL, SpringerLink, and Science Direct), we add them as 'referenced' for those databases, and if they are from other sources, we add them to the 'other' section. Table 1 shows the results of our search queries. We extensively investigated the relevant works once we had reduced them to ninety-four publications to uncover current trends in fairness issues, adopted methodologies to solve these issues, fairness-related must-know terminologies, remaining challenges, popularly utilized datasets, and developed tools in this area. We also discussed the probable scope of improvement in model fairness disclosed in some of those filtered articles and from our understanding.

4 Result Of The Mapping Study

Scholars have devoted considerable attention to exploring the counterfactual concept in machine learning and artificial intelligence to ensure fair prediction. In our study, we searched 420 research articles to identify contributions in this field, ultimately selecting 94 articles that closely aligned with the scope of fairness. In the following subsections, we represented our findings by answering the first two research questions enumerated in Section 3.1.

4.1 State of the art of research in Ethical AI

Although researchers have been studying machine learning models since the early nineteenth century, the unfairness of predictive machine learning models is a relatively recent topic.

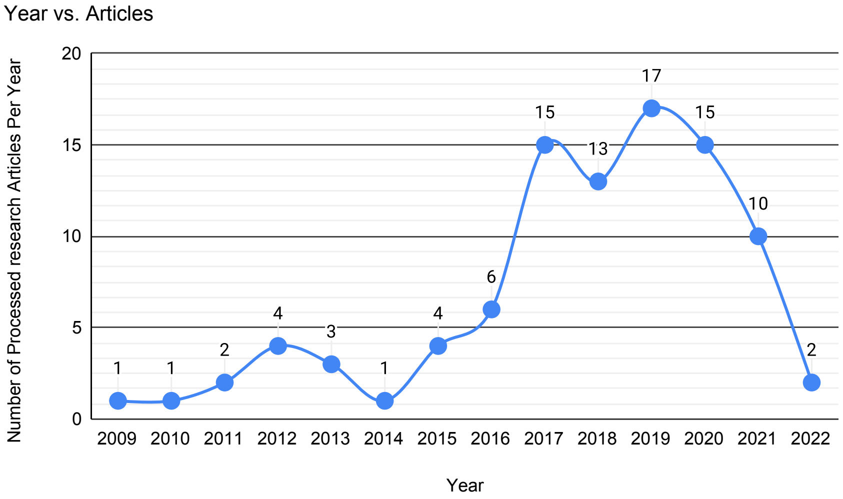

Fig. 4 Number of papers found per year starting from 2009 to 2022.

Fig. 4 displays the number of processed papers per year, revealing a significant increase in the number of papers after year 2016. It indicates the growing interest of academics in the field of fair prediction.

4.2 Engaged Countries and Individuals

Fig. 5 indicates that concerns regarding fairness in ML and AI models have gained widespread attention and are not limited to any specific group of researchers. During our analysis, we did not notice any particular author with significantly more publications. However, we noticed that many articles came from authors from the United States. Out of the 94 papers analyzed, the highest number of publications on this topic came from the United States, followed by the UK with approximately one-fifth of the author's numbers of the US, and Germany in third place with roughly one-seventh of the author's numbers of the US.

5 Addressed Fairness Issues

The filtered articles describe some challenges, and we summarized them into a few common problem groups: biased training data, bias toward feature groups, biased decision models, lack of prediction transparency, and inherent bias. We discuss some of the common key factors that cause these biases, represented in a graphical view in Fig. 6. Bold oval shapes represent the issues, and the other oval shapes, with multiple outward arrows, represent the key factors. The arrows represent the contribution of key factors in generalized fairness issues.

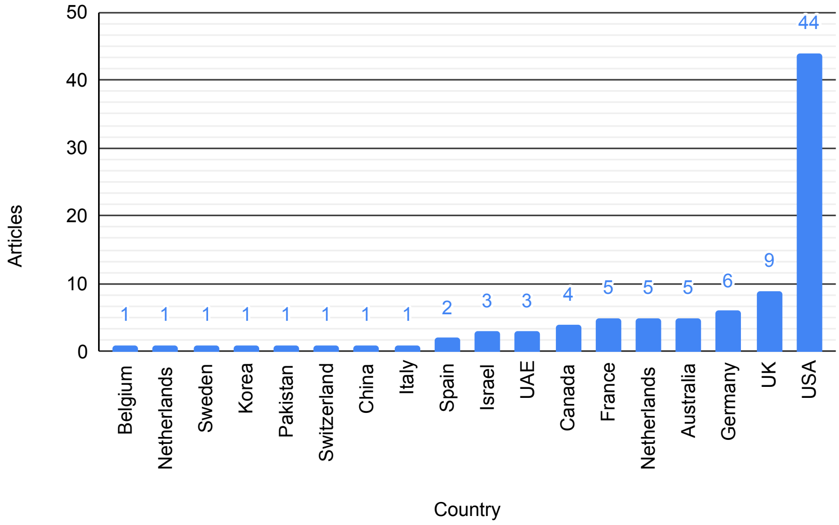

Fig. 5 Number of articles in processed papers for this mapping study per country.

5.1 Biased training data

Bias in the data refers to the presence of systematic errors or inaccuracies that deplete the fairness of a model if we use these biased data to train a model. Bias can potentially exist in all data types as bias can arise from a list of factors [95]. Among these factors, **measurement bias** is a potential source of data bias in ML that occurs when the measurements or assessments used to collect data systematically overestimate or underestimate the true value of the characteristic being measured [85, 103, 110, 111]. For example, training data in criminal justice systems often includes prior arrests and family/friend arrests as attributes to assess the probability of repeating a crime in the future. As a result, it can lead to racial profiling or disparities in sentencing practices because we cannot confidently guarantee that an individual from a group will behave similarly to others.

Next, **representation bias** is another crucial factor for a biased training dataset. It refers to the bias in a dataset or model that results from under or over-representing certain groups or characteristics in the data, which can lead to biased or inaccurate predictions for those groups [91, 103, 104]. For example, Yang et al. have highlighted the issue of underrepresented images, particularly for the representation of people and their attributes in the ImageNet dataset, where only a small percentage of images (1%, 2.1%) are from China and India, and a comparatively more extensive portion of images are from the United States (45%) [91]. Similarly, Shankar et al. demonstrated that classifier performance is notably lower for underrepresented categories trained on ImageNet [104]. Additionally, word embeddings, learned from large corpora of text data, represent words as vectors in a high-dimensional space in various NLP (Natural Language Processing)-based predictive models. However, these word embeddings encode various issues, such as gender biases and lack of diversity, which contribute to

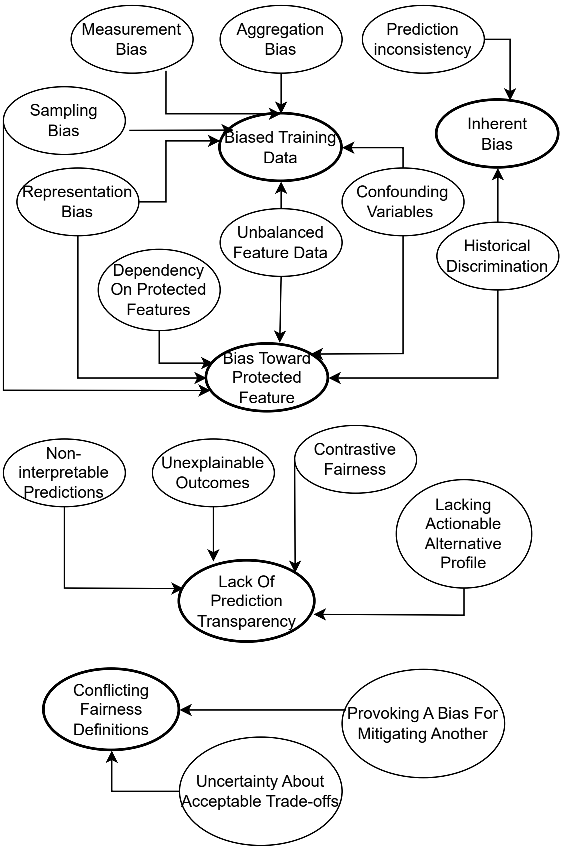
Measurement Bias
Aggregation Bias
Prediction inconsistency
Sampling Bias
Biased Training Data
Inherent Bias
Representation Bias
Confounding Variables
Unbalanced Feature Data
Historical Discrimination
Dependency On Protected Features
Bias Toward Protected Feature
Non-interpretable Predictions
Unexplainable Outcomes
Contrastive Fairness
Lacking Actionable Alternative Profile
Lack Of Prediction Transparency
Conflicting Fairness Definitions
Provoking A Bias For Mitigating Another
Uncertainty About Acceptable Trade-offs

Fig. 6 Key factors contributing to training data bias, bias toward protected features, lack of prediction transparency, and conflicting fairness definitions.

the model's inability to generalize well to new data [86, 87, 123]. Also, when we train a word embedding model on a dataset where the word "doctor" is more associated with the word "man" and "nurse" is more associated with "woman", the model may learn to associate gender with these professions, even when it is not necessary for the given task. In this regard, Zhang et al. argue that while researchers designed many machine learning models to optimize and maximize accuracy, they may also inadvertently learn and propagate existing biases in the training data [92].

Sampling bias slightly differs from the representation bias [103, 131]. Sampling bias occurs when the sample data for training does not represent the population targeted to generalize. In contrast, representation bias is an inadequate representation of the real-world distribution of the data. For example, if a researcher wants to study the height of people in a particular country but only samples people from a single city, the results may only represent part of the country's population. The sample may be biased toward people from that specific city, resulting in inaccurate conclusions about the height of the country's population. Subsequently, another aspect that can make the model predictions inaccurate is **label bias** [92]. It occurs when the labels assigned to data instances are biased in some way. For example, a dataset of movie reviews may have been labeled by individuals with a particular preference for a certain genre, leading to biased labels for movies of other genres.

Besides them, **Aggregation bias** refers to a type of bias that arises when a model is used to make predictions or decisions for groups of individuals with different characteristics or from different populations [113]. It occurs when a single model is used to generalize across different groups or sub-populations and can lead to sub-optimal performance for some groups. For example, scholars study blood glucose (sugar) levels such as HbA1c (widely used to diagnose and monitor diabetes), which usually differ across ethnicities and genders. Thus, a single model may become biased towards the dominant population and not work equally well for all groups (if combined with representation bias) [61].

Depending on the specific application and context, there may also be other sources of bias in training data that can potentially lead to unfair model outcomes. Thus, processing the training data to remove existing bias is often necessary for many machine learning-based models [92]. Otherwise, If we train the model on a biased dataset, it can lead to unfair outcomes and perpetuate existing societal inequalities. Therefore, it is essential to identify and mitigate bias in the data to ensure that machine learning models are fair and equitable. Scholars in the articles primarily address this step as pre-processing [123].

5.2 Inherent bias

Inherent bias, also known as intrinsic bias, refers to the bias inherent in the studied data or problem rather than the bias introduced during the modeling or analysis process [62]. Along with all the discussed biases, we can observe inherent biases in multiple ways, such as prediction inconsistency and prediction falsification due to partial data. **Prediction inconsistency** is a different type of bias addressed as leave-one-out unfairness. Although a definite cause is yet to be discovered, scholars often held many of the above biases responsible for prediction inconsistency[84]. Prediction, regardless of its accuracy, is expected to be constant. This bias refers to a situation where including or removing only one instance from the dataset and retraining the model on this modified dataset alters the prediction outcome for another instance entirely irrelevant to that included or deleted instance [84]. In other words, the model's predictions are not consistently fair for all individuals in the dataset when the model is retrained on the remaining data after removing a single data point. As a result, the predictions for the removed data point may change in an unfair or biased way. The leave-one-out unfairness problem is particularly relevant for datasets where individual data points are sensitive. It makes ML models unreliable and untrustworthy in serious implementations such as predicting recidivism or determining creditworthiness criminal detection.

Another significant inherent bias source is the **Historical discrimination**. Even if the algorithms used in decision-making processes are unbiased, the data they are trained on may contain historical biases, leading to discriminatory outcomes [62]. Calmon et al. stated that the presence of historical discrimination in linear datasets, such as the widely studied Adult Income dataset used for evaluating fairness in machine learning, can result in biased predictions despite high model performance when using such biased data for training [97, 124]. For example, suppose a training dataset for an employee hiring algorithm only includes data from past hires, and past hiring practices were biased against certain groups. In that case, the algorithm will continue perpetuating that bias in hiring decisions. Historical bias can be challenging to address because it reflects broader societal biases deeply ingrained in our institutions and culture. Even if we design a fair decision-making system according to a particular definition of fairness, the data it uses to learn may still reflect historical biases and lead to unfair decisions [105]. However, it is vital to recognize and address historical bias in machine learning models to prevent perpetuating unfair and discriminatory practices.

5.3 Bias toward feature groups

Bias towards protected feature groups refers to a type of bias where a machine learning algorithm may unintentionally favor or discriminate against certain protected feature groups, such as women, colored, or ethnic minorities when making decisions. This bias is problematic because it perpetuates existing societal inequalities and can result in unfair outcomes for specific individuals or groups, especially in high-stakes domains [98, 125]. For example, a protected feature-dependent model for predicting recidivism rates could result in more false positives for certain groups, leading to lengthy prison sentences or increased monitoring, even when the individual may not pose a significant risk. Many factors can lead to bias toward protected feature groups. For example, training and inherent data bias can also be responsible for discriminating against people with protected feature groups. Also, other factors, such as unbalanced feature data, confounding variables, and predicting attribute's connection to protected feature can contribute to this bias in addition to those mentioned in section 5.1.

We refer to a dataset with severely skewed or uneven value distribution across various features as having **unbalanced feature data**. In other words, when a dataset has a significantly larger or smaller number of instances of certain features or categories within features compared to others, it indicates unbalanced feature data. For example, suppose we utilize a model that announces verdicts, and the training data contains gender information as a data feature. If, in the data, females are verdict more times than males for training an RAI, the RAI model may perpetuate these biases and unfairly target females (specific groups) [67].

Confounding variables can be another reason the model is biased toward certain feature groups [114]. If a protected feature correlates with other variables that affect the outcome variable, then the protected feature can become a confounding variable. It can lead to training a model in a biased way that can associate the confounding variables with the outcome variable. For example, consider a study investigating the relationship between caffeine consumption and heart disease risk. If the study does not control for age, then age may act as a confounding variable, as older people tend to have a higher risk of heart disease and may also consume more caffeine. In this case, the study may mistakenly conclude that caffeine consumption is associated with a higher risk of heart disease when in fact, it is the confounding variable of age that is responsible for the observed relationship.

Additionally, **dependency on protected features** may lead to poorer outcomes [98]. When a machine learning model relies heavily on protected features, it can lead to biased predictions that favor certain protected groups over others. For example, a loan approval model that relies heavily on race as a feature may be biased against certain racial groups. It may happen if the model fails to identify other strongly correlated features that are not sensitive or if the dataset lacks enough features other than the protected feature. As a result, the model may unfairly deny loans to members of certain groups.

Other than these reasons, scholars also mention other aspects, such as Dwork et al. stating that the inability to learn the distribution of sensitive attributes in the training data is a potential reason for bias towards protected features [99]. The authors define `sensitive attributes` as those protected by anti-discrimination laws (race, gender, and age). Mishler et al. discussed that if we train RAI models on datasets having sensitive features, they may become biased against certain races or gender [67]. As the reason for bias toward feature groups, some articles also claim that false positive outputs are as harmful as false negative outputs in many high-stakes decisions for a dataset with protected attributes [63]. For example, in a criminal justice system, falsely predicting someone is likely to re-offend (a false positive) could lead to unjust incarceration or other forms of harm [64, 115, 117].

5.4 Decision model bias

The above sections 5.1, 5.2, and 5.3 describe how data bias can ruin fair predictions for some ML models. However, a predictive ML model can be unfair even though the training dataset is not biased or contains protected attributes such as race, gender, or age [98, 125, 132]. **Algorithmic bias** is a potential bias that can introduce discrimination or unfairness in the model. It refers to the bias introduced by the algorithm rather than inherent in the input data [88, 118].

Another reason is the **hidden biases**. Despite having a balanced distribution in the dataset and being free of sensitive feature correlation, it still can contain hidden biases, which refer to the biases present in the data used to train an ML model that is not immediately apparent or identifiable [45]. These biases can result in discriminatory outcomes for specific groups. For example, using zip codes in the model may inadvertently incorporate racial or economic factors that are not directly related to criminal behavior. Using zip code as an attribute can lead to over-predicting the likelihood of recidivism for specific groups and under-predicting it for others, resulting in unjust outcomes.

Besides them, many Risk Assessment Instruments (RAI) implement ML-based models and may only **emphasize prediction accuracy**, which can eventually lead to unfairness [132]. Risk assessment instruments (RAIs) are machine learning models used to assess the likelihood of recidivism or future criminal behavior in individuals [110, 111, 119]. Unfairness in these tools can lead to severe consequences that are not adequately justified, such as serving more years in jail for being colored individuals [110, 111].

Additionally, **evaluation bias** refers to a type of bias that arises while evaluating machine learning models, and thus, it is not related to data bias. It happens when the performance of a model is assessed in a way that is biased toward certain groups or outcomes, leading to misleading or incorrect conclusions [95, 103]. For example, suppose we adopt an evaluation method solely based on its overall accuracy without considering the model's performance on different subgroups. In that case, the evaluation outcome may hide that the model performs poorly on certain protected groups while delivering high accuracy. It can lead to adopting biased models that appear to perform well overall but are discriminatory towards certain groups. Mitigating these biases can help ensure a fair model, build trust in machine learning systems, and increase their adoption in various domains.

5.5 Lack of prediction transparency

Machine learning models can be complex and challenging to interpret, making it hard to understand how the model makes decisions and identify potential sources of bias [89, 90, 106, 120]. A model can be unfair if a model lacks transparency. Authors identified transparency issues generated while developing the ML algorithm, including non-interpretable predictions [59], unexplainable outcomes [58], lack of contrastive fairness [42], lack of transparency [73] and lack of actionable alternative profiles [70]. These issues can lead to unexpected vulnerabilities, hidden biases, and negative impacts on various stakeholders [58, 68–70, 81].

First, **Non-interpretable predictions** of ML models refer to predictions made by models that humans need help to understand meaningfully. Here, non-interpretable predictions can occur when the model is very complex, such as deep neural networks and adversarial networks, or when the training data is too large or diverse to be easily understood, for example, latent representations of an encoder [59]. Next, **unexplainable outcomes** in machine learning refers to situations where the model's predictions need more justification [58]. The machine learning model provides a single outcome without explaining why the model picked this particular choice out of several possibilities in the final forecast. It makes it difficult for humans to understand how and why the model arrived at a particular decision [48, 50, 52, 57].

Moreover, contrastive fairness aims to ensure fairness in decisions by comparing outcomes for similar individuals who differ only in a protected attribute (such as race or gender). **Lack of contrastive fairness** in models can make the model biased favorably or unfavorably towards a group of stakeholders [42]. For example, if a job screening model is biased toward male candidates over females with similar qualifications, the company must modify the algorithm to consider them equally. Lastly, **lack of actionable alternative profiles** limits the model's capability to produce other feature value combinations that would help to generate an expected output. Actionable alternative profile refers to providing a set of alternative actions or decisions that could be taken in response to the outcome of a machine learning model [70]. For example, a machine learning model in medical diagnosis may predict a patient's high risk of developing a particular disease. However, instead of just providing this information to the healthcare provider, the model could also suggest alternative courses of action or treatment options that could reduce the risk or prevent the disease. Having actionable alternative profiles is crucial for ensuring the reliability of a decision, as more than relying on a single decision may be required.

5.6 Multiple definitions of fairness

Many scholars propose different definitions of model fairness from various perspectives to address these issues. Different definitions of fairness often lead to conflicting objectives, challenging developers and policymakers. For instance, group fairness requires equal treatment of different protected groups, while individual fairness demands that the model treat similar individuals similarly. Ensuring equal outcomes for all protected groups may require setting different thresholds for different groups, which may violate the principle of treating individuals equally regardless of their group membership [93]. Conversely, ensuring equal treatment for all individuals may lead to unequal outcomes, which may be unfair. Meanwhile, predictive parity focuses on equalizing the proportion of true positives across different groups. In contrast, equalized odds aim to balance `true positive` and `false positive` rates for a model's prediction.

In practice, implementing one definition of fairness may cause violations of other definitions, leading to a trade-off between competing objectives. Also, if a model is designed to be fair according to a particular definition of fairness, it may still exhibit unintended biases and unfairness when used in practice. Therefore, it is essential to consider multiple definitions of fairness and the trade-offs between them when designing and evaluating machine learning models to minimize the risk of creating discriminatory outcomes. Therefore, there is often a trade-off between different notions of fairness that the model must carefully consider for decision-making systems. A few articles discuss the challenges of defining and achieving variously defined fairness in machine learning models and propose various solutions to address these challenges [98, 99, 105].

6 Adopted methodologies by authors to solve these issues

Authors from the filtered papers adopted many techniques to solve these biases. While some authors have proposed practical models to remove and detect data bias, protected-feature discrimination, unexplainable prediction, model bias, or prediction inconsistency, some authors have provided fairness definitions and metrics to remove ambiguity and discuss the trade-offs of these concepts. We generalize and classify these methodologies according to the specific problem types they solve. Fig. 7 depicts the methodologies scholars have followed to solve generalized issues. The figure's first column (containing 'Biased training data,' 'inherent bias,' 'bias toward protected feature group,' 'decision model bias,' 'lack of prediction transparency,' and 'multiple definitions of fairness') contains all the generalized issues. We also discuss these classes elaborately in the following sections.

6.1 Methodologies to mitigate data bias

We generalize the techniques adopted by the authors of the filtered articles.

6.1.1 Extending and diversifying

In this technique, people modify the data to diversify the model's input data and implement it for identifying bias and modifying the model [96, 121, 122, 129, 133].

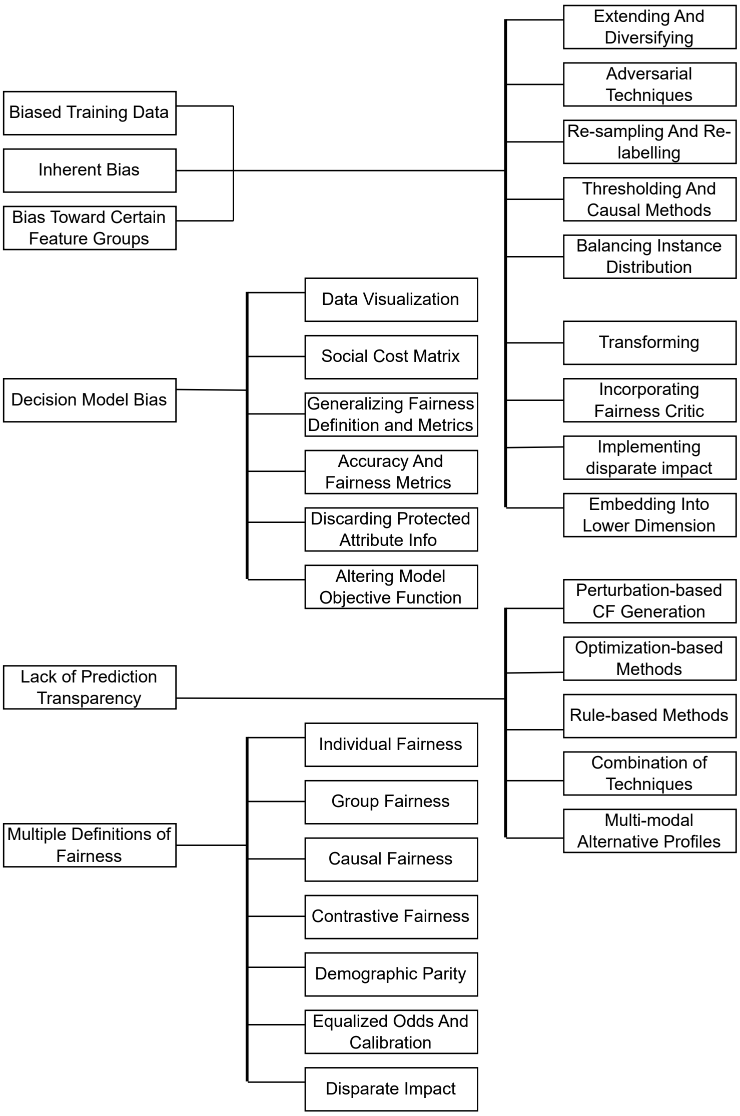
Biased Training Data
Inherent Bias
Bias Toward Certain Feature Groups
Extending And Diversifying
Adversarial Techniques
Re-sampling And Re-labelling
Thresholding And Causal Methods
Balancing Instance Distribution
Transforming
Incorporating Fairness Critic
Implementing disparate impact
Embedding Into Lower Dimension
Decision Model Bias
Data Visualization
Social Cost Matrix
Generalizing Fairness Definition and Metrics
Accuracy And Fairness Metrics
Discarding Protected Attribute Info
Altering Model Objective Function
Lack of Prediction Transparency
Perturbation-based CF Generation
Optimization-based Methods
Rule-based Methods
Combination of Techniques
Multi-modal Alternative Profiles
Multiple Definitions of Fairness
Individual Fairness
Group Fairness
Causal Fairness
Contrastive Fairness
Demographic Parity
Equalized Odds And Calibration
Disparate Impact

Fig. 7 Classification of adopted methodologies to solve several types of issues.

One method proposes an approach to understanding a model's bias sources by adding counterfactual instances in the data points. First, they propose to modify the input data to create diversified new data points similar to the original data points but with more critical features changed. Then, the model identifies and quantifies any bias in the model by comparing the model's predictions on the original data points and the corresponding counterfactual instances [121]. Another method follows this procedure, uses path-specific counterfactuals, and adjusts for bias along specific paths [129]. Similar to these techniques, the ‘Counterfactual Fairness with Regularization (CFR)’ method aims to remove the direct effect of sensitive attributes on the predicted outcome while preserving as much accuracy as possible. The approach involves constructing counterfactual instances and ensuring fairness for each individual under different sensitive attribute values and then using regularization to encourage the model to make similar predictions for similar individuals with different sensitive attribute values [133]. This method ensures individual fairness, and there are other fairness concepts similar to counterfactuals, such as the group fairness assumption and the counterfactual fairness assumption. Some scholars also propose integrating all these counterfactual fairness concepts into the model similarly for unbiased classification, clustering, and regression [122].

6.1.2 Adversarial techniques

Many scholars discuss adversarial techniques for data debiasing by removing or reducing the impact of sensitive features that could lead to biased predictions [98, 116]. In adversarial training, we generally train a model to predict the outcome while being attacked by an adversary trying to infer the sensitive attributes. One adversarial technique involves training a neural network with a different fairness branch to prevent bias based on a protected attribute in the learned representations. They train the network adversarially, where the fairness branch competes against the main classification task to achieve accuracy and fairness [28]. Zhang et al. proposed an adversarial technique with a predicting branch/network and an adversary branch/network [92]. The primary or predictor branch predicts Y given X and learns the weight W with an optimization function like stochastic gradient descent (SGD) aiming to minimize the loss. The output layer passes through the adversary branch aiming to predict Z. The architecture of the adversary network depends on the fairness issue they aim to solve. In aiming for goals like *Demographic Parity*, the adversary would predict the protected variable Z using only the input’s predicted label $\hat{Y}$ (and not the actual label Y), while withholding its own learning process. Similarly, for achieving *Equality of Odds*, the adversary utilizes both the predicted label $\hat{Y}$ and the true label Y. For achieving *Equality of Opportunity* for a specific class y, the adversary restricts its instances to only those where $Y = y$ [31]. *Demographic Parity*, *Equality of Odds* and *Equality of Opportunity* is defined in section 6.5. Fig. 8 depicts a visualization of the architecture for such adversarial techniques.

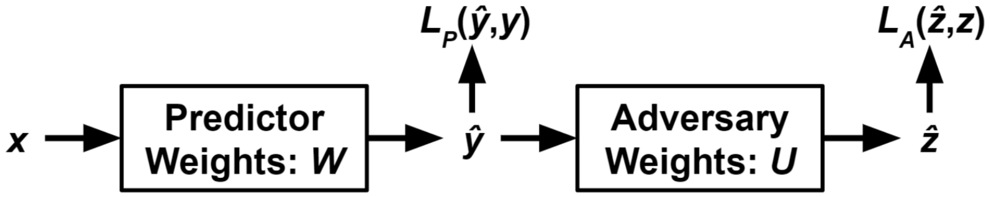

Fig. 8 The architecture of adversarial network visualized by Zhang et al. [92].

Some scholars explore GAN, where the generator generates indistinguishable synthetic data while the discriminator differentiates between real-world and synthetic data. To ensure fairness, a fairness critic in an additional adversarial objective ensures that the representations learned by the generator are not biased toward any particular group [46]. Besides them, another method divides the training into two deep neural networks: a representation network predicts the protected attribute and another network keeps training on the source dataset and fine-tuning on a target dataset to maintain high accuracy and transferability across datasets [100].

6.1.3 Re-sampling and Re-labelling

Imbalance in some dataset features contributes to developing data bias [107]. As a result, some researchers explore pre-processing the dataset to mitigate dataset bias. Re-sampling and re-labeling are two such processes, and many research results validated their effectiveness. Re-sampling addresses data imbalance that causes bias in machine learning models. In a dataset, if the number of instances belonging to one class is significantly higher than the other classes, then the model may be biased towards the majority class. Re-sampling techniques refer to oversampling the minority class or undersampling the majority class to create a balanced dataset. It ensures more representative data, diverse data from various sources and populations, and balanced data across different groups [92, 98]. Besides re-sampling the input data, scholars also propose re-labeling data instances to mitigate bias. Bolukbasi et al. proposed to modify the training data by explicitly identifying gender-neutral words and using them to adjust the gender-specific words in the embedding data [123]. For example, they replace the word "he" with "she" and vice versa in the text data, creating balanced examples of each gender association. They also use gender-neutral word pairs (no association with a particular gender), such as "doctor" and "nurse", to help the model learn a more balanced representation of gender-related concepts [123]. In this regard, Kamiran et al. proposed a 'massaging' method that used and extended a Naïve Bayesian classifier to rank and learn the best candidates for re-labeling [26, 63].

6.1.4 Thresholding and Causal methods

Researchers also explore post-processing techniques, such as thresholding for training data bias removal, especially for RAIs [25, 27, 134]. In the post-processing technique, scholars adjust the predictions to meet a fairness constraint after training the model on data. Thresholding is one such post-processing technique where a threshold is set on the model's output to ensure a certain level of fairness. For example, Hardt et al. introduce the concept of "equality of opportunity", which means that each group's `true positive rate` and `false positive rate` should be equal, regardless of the protected attribute [125]. They propose a method to enforce this constraint using a constrained optimization problem, penalizing models with disparate impacts on different groups while maximizing overall accuracy. Other than thresholding, causal methods can ensure model fairness by analyzing and modeling the causal relationships between input features, the predicted outcome, and the sensitive attribute. The goal is to identify the direct and indirect causal relationships between these variables and to use this information to create fair and unbiased models. In this context, Salimi et al. propose that repairing the training data to remove unfair causal relationships is more efficient than implementing correlation-based fairness metrics to address the root causes of unfairness. They propose a framework with three steps: identifying the causal relationships between the input features and the output label, identifying any causal relationships between the input features and protected attributes, and then implementing causal inference and database repair techniques to remove any unfair causal relationships between the input features and the output label [94]. Like these techniques, Razieh et al. use causal inference methods to estimate counterfactual outcomes for different subgroups and then use these estimates to make fair predictions [135]. Furthermore, Depeng et al. propose the GAN that they train with a novel loss function that penalizes the model for violating causal constraints. The proposed method ensures that the model does not use the protected attribute to make predictions, thereby reducing bias in the data [112].

6.1.5 Balancing instance distribution

Lastly, With the goal of accurate image classification models, Yang et al. introduce a two-step approach to filtering and balancing the distribution of images in the popular Imagenet dataset of people from different subgroups [91]. In the filtering step, they remove inappropriate images that reinforce harmful stereotypes or depict people in degrading ways. In the balancing step, they adjust the distribution of images from different racial and gender subgroups to ensure that the dataset represents each subgroup equally. They supported their proposal by showing fairer classification performance than in classification in the original Imagenet dataset [91]. Along with distribution balancing, some articles also propose data collection processes to determine if the predictions require fairness, inspiring model modification to ensure fairness. For example, Buolamwini et al. propose the Gender Shades Benchmark, a benchmark dataset designed to evaluate gender classification systems regarding intersectional accuracy disparities. This dataset consists of a diverse and representative set of images that includes darker-skinned individuals and women who do not conform to traditional gender norms [95].

6.2 Methodologies to mitigate bias towards protected features

To mitigate bias toward certain groups, scholars propose identifying the source of the bias first and then mitigating the bias along the route. Many of the methods mentioned above maintain this trend, whereas some filtered articles specifically explore protected feature bias mitigation strategies more. In addition to the methods mentioned earlier, scholars also consider additional techniques to reduce bias against protected attributes.

6.2.1 Implementing transformation theory

Transformation theory is a framework for improving model fairness by transforming the input data to mitigate the effect of sensitive attributes on the model's predictions. The model first predicts the protected attribute and then uses this to generate transformed data that removes the effect of the sensitive attribute. Then, they generate a fair model by training in the transformed data. For example, Meike et al. map the input data into a metric space where distances represent the similarity between individuals and find the minimum-cost way to transport the distribution of protected attribute values from biased to unbiased dataset [65]. Paula et al. followed this approach and proposed two methods, one that uses a generative adversarial network to learn the optimal transport plan and another that directly estimates the transport plan using a convex optimization algorithm. They expect the resulting model to achieve fairness for the protected attribute while maintaining accuracy [101]. Some scholars also have explored convex objective functions to minimize the correlation in previous years [124].

6.2.2 Implementing a fairness critic

To understand if a model is biased toward certain groups, scholars explored and proposed several theoretical and practical fairness concepts to mitigate bias towards protected groups. Implementing a fairness critic network to learn a fair representation of the data by training an adversarial network is another remarkable idea to mitigate bias against protected attributes. Researchers mainly implement this idea for RAI models (classifier models). The fairness critic is a separate neural network that takes the learned representations as input and outputs a score that reflects the level of bias in the representation. The classifier maximizes prediction accuracy, while the fairness critic maximizes fairness. The two objectives compete against each other in a minimax game. By doing so, the model learns a representation that maximizes accuracy while minimizing bias towards protected attributes [108]. Scholars also approach generative approaches combined with a fairness critic network. For example, Depeng et al. introduced an adversarial objective-based fairness loss function in the GAN framework to generate realistic and unbiased, specifically concerning protected attributes [136]. The critic network distinguishes between valid and generated (that violate fairness constraints) samples. The model encompasses a generator G_{Dec} with a conditional distribution $P_G(x, y, s)$ that generates the fake data.

$$P_G(x, y, s) = P_G(x, y|s)P_G(s) \tag{1}$$

In equation (1) x, y = data pair, s = protected or sensitive attribute, $P_G(s) = P_{data}(s)$, $P_G(x, y|s = 1) = P_G(x, y|s = 0)$ for ensuring statistical parity constraints and the fairness critic differentiates the two types of generated samples $P_G(x, y|s = 1)$ and $P_G(x, y|s = 0)$ indicating if the synthetic samples are from protected or unprotected groups.

Fig. 9 represents a visualization of their model [136]. The proposed approach is evaluated on several benchmark datasets and shown to produce realistic and fair samples [136].

6.2.3 Disparate impact-based methods

Recently, scholars also explored the disparate impact of recidivism prediction instruments and offered several solutions to ensure protected group fairness in these tools [64, 93, 115, 124]. "disparate impact remover" is one such method. It modifies the data distribution to ensure equal group representation in the training data [124]. The method achieves this by adjusting the weights of each training example based

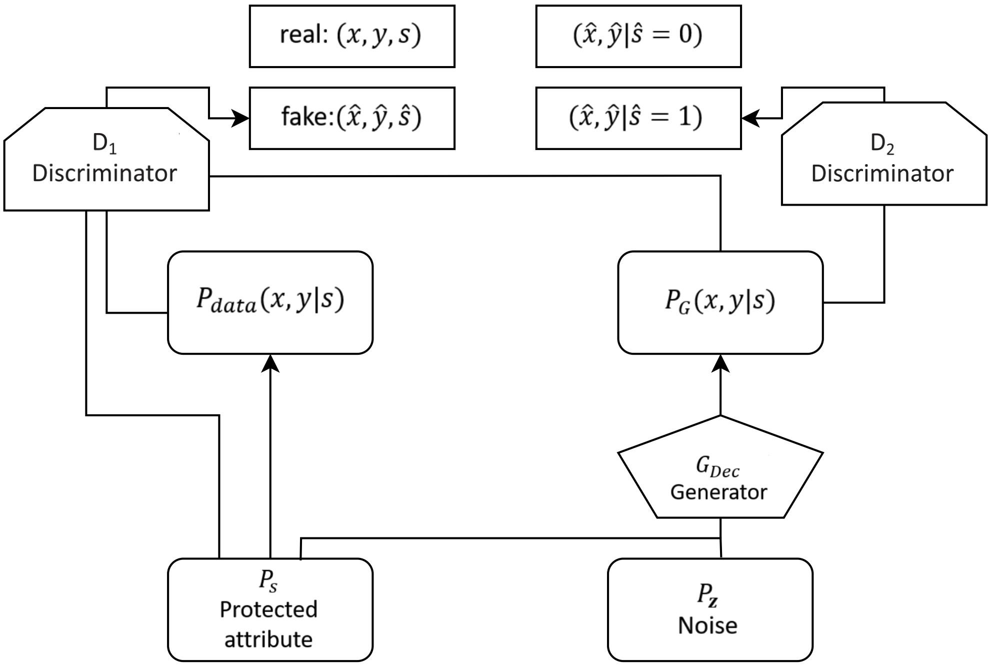
real: (x, y, s)
fake: $(\hat{x}, \hat{y}, \hat{s})$
$(\hat{x}, \hat{y}|\hat{s} = 0)$
$(\hat{x}, \hat{y}|\hat{s} = 1)$
D_1 Discriminator
D_2 Discriminator
$P_{data}(x, y|s)$
$P_G(x, y|s)$
G_{Dec} Generator
P_s Protected attribute
P_z Noise

Fig. 9 Structure of FairGAN as represented in [136].

on the protected attribute's distribution, thus equalizing the acceptance rate across groups. Also, there is a framework for evaluating the fairness of prediction models and demonstrating how to apply it to assess the fairness of recidivism prediction instruments [64]. This framework has three steps: identifying the group protected by anti-discrimination laws or ethical considerations, identifying the outcome variable, and evaluating prediction fairness through the "disparate impact" and "equal opportunity" tests.

6.2.4 Embedding data into lower-dimension

Another renowned protected-feature bias mitigation approach is to learn fair representations by mapping the original data into a lower-dimensional space that preserves the relevant information for the downstream task while removing the discriminatory features [98]. This method aims to learn a representation invariant to protected attributes such as race or gender, thereby ensuring that the downstream classifier will not make decisions based on these attributes.

6.3 Methodologies to mitigate model bias

6.3.1 Visualizing data

Some articles explore visualizing the data effectively to identify the data source and remove model bias, as hidden biases contribute to discrimination in ML-model prediction. For example, Dwork et al. propose an interactive visualization tool called the "What-if tool" to increase awareness of the potential sources of discrimination in machine learning models [99]. It provides an intuitive and user-friendly interface that allows users to explore the impact of various changes on the fairness and accuracy of machine learning models in real-time and visualize its outputs. Users can change input data points, adjust thresholds, and modify other parameters to observe how different decisions affect model performance and fairness through a variety of fairness and performance metrics.

6.3.2 Introducing social cost matrix into misclassification cost matrix

Developers also implement ML models, especially classification models, as risk assessment instruments (RAI). Researchers proposed several approaches to eliminate RAI bias, such as CF-based and adversarial debiasing-based approaches. Firstly, Kamiran et al. proposed a CF-based solution using a decision-theoretic framework in classification [44]. Their proposed framework extends the standard cost-sensitive learning approach by introducing a social cost matrix to capture the societal costs associated with different types of errors. The model also includes a two-step process for building a discrimination-aware classifier. In the first step, the model learns a classifier that minimizes the expected cost of misclassification. In the second step, the model modifies the cost matrix with the social cost matrix to compute the optimal decision boundary using the modified cost matrix [44]. They further demonstrated the implementation of the discrimination-aware classifier using a threshold-based approach. The threshold-based approach uses the expected costs of misclassification and discrimination to compute a threshold value that balances the trade-off between the two costs. The classifier then classifies instances as positive or negative based on whether their predicted probability exceeds the threshold value [44].

6.3.3 Generalizing fairness definition and metrics

Definition development for fairness terminologies and metrics for measuring fairness in the model outcome is necessary before developing fair models and bias reduction techniques. Thus, many researchers have attempted to propose fairness-related terminologies by generalizing definitions of fairness from psychology, statistics, quantum computing, and many more fields. Scholars implement these variously proposed fairness concepts in a model development step. For example, after defining "equalized odds" and "calibration", Kleinberg et al. propose a statistical method called the "direct constrained optimization" method. They formulate the model optimization problem as a constrained optimization problem to maximize accuracy subject to fairness constraints. This optimization process involves solving for a set of model parameters that satisfy the fairness constraints, such as requiring equal false positive or false negative rates across groups or limiting the difference in average risk scores between groups while maximizing the model's accuracy. They further adjust the risk scores by including a post-processing step [132]. Lum et al. propose another statistical approach similar to the previous one by incorporating "group fairness" for calculating the trade-off and solving the constrained optimization problem using convex optimization [109]. A few articles also propose new metrics such as calibration error to quantify the performance of calibrations of these methods[126]. A calibration error metric measures the discrepancy between predicted probabilities and the observed outcomes in an ML model. It quantifies the calibration performance of the model by evaluating how well the model predicted the probabilities and how they align with the true probabilities of the events or classes. Afterward, "disparate impact-aware Naive Bayes" and "equalized odds-aware Naive Bayes" based strategies are two noteworthy fairness approaches that ensure that the predictions are disparate impact-free and the false positive and negative rates are equal across different protected groups [66]. We further discuss all the proposed fairness-related definitions of the filtered papers in section 6.5.

6.3.4 Incorporating accuracy and fairness metrics

Usually, any ML models follow accuracy matrices for developing the model. However, to decrease prediction bias, some scholars propose statistical methods incorporating fairness and accuracy metrics while developing an ML model and balancing the trade-offs between fairness and accuracy, especially in risk assessment instruments [109, 126, 132].

6.3.5 Protected attribute info discarding

The adversarial debiasing method tries to learn a debiased representation of the data by training a neural network to predict an outcome while at the same time being forced to discard any information about the protected attribute. Madras et al. propose an adversarial training-based method to address issues of fairness and bias in machine learning models [100]. The discriminator predicts the sensitive attribute from the learned representation, while the generator produces a representation that is both predictive of the task and fair. Another approach is the "jointly constrained Naive Bayes", which restricts the classifier to use only a subset of features that are minimally correlated with the protected attribute [66]. Furthermore, Calmon et al. proposed quantifying the relationship between protected attributes and other features and minimizing the mutual information between protected and remaining features. Finally, the authors stated that the sensitive attribute is not used to make classification decisions and presented convincing experimental results to support that fairness and accuracy are balanced [47].

6.3.6 Statistical framework to alter objective function

The paper proposes a statistical framework for developing fair predictive algorithms that explicitly consider fairness constraints during model training. It introduces a fairness penalty term to the objective function that penalizes the algorithm for its deviation from a desired level of fairness. This framework is designed to balance fairness and accuracy and can be applied to a range of machine learning models [109].

6.4 Methodologies to ensure prediction transparency, explainability, interpretability

Researchers have emphasized providing prediction explanations and interpretations to maintain transparency of the model predictions. Explanation and interpretation of the prediction confirm the outcome's legitimacy and transparency about fairness [59, 79]. In addition, prediction explanation provides alternative feature combinations to modify the model outcome, and prediction interpretation validates the model's outcome (probability of other outcome occurrences without dependency on protected attributes) [50, 68, 72]. To explain and interpret a model's outcome from the fairness perspective, Counterfactual explanations have gained significant popularity among the various approaches [71]. The counterfactual analysis involves asking "what-if" questions to determine how changing one or more features of a particular instance would affect the model's output. We can use this technique to identify instances where a model's output may be unfair and to make corrections to improve fairness [58]. Many scholars have proposed counterfactual approaches that received scholars' attention, such as performance improvement for multi-agent learning, causal inference in machine learning, explanation for system decision of black-box models, the actionable alternative outcome of existing and new AI models, etc. [70, 137–142]. Academics are exploring trade-offs to generate counterfactual explanations and methods to utilize the generated CFs to provide explainable and interpretable model outcomes.

6.4.1 Perturbation-based methods

One common approach to generating CF is the perturbation-based method. A perturbation-based method is used in various fields, including machine learning and optimization, to analyze the behavior of a system by modifying input features and keeping the remaining features fixed to observe changes in the prediction. In machine learning, scholars mainly employ perturbation-based methods to assess a model's robustness, sensitivity, or generalization. Perturbation distance, or feature or input distance, measures the extent of modification or change applied to input features when generating counterfactual explanations. Wachter et al. proposed a method to generate CF explanations for predictions without accessing the model's internal architecture. This method turns the CF generation problem into an optimization problem with an objective function that considers measuring perturbation distance between instances. Search algorithms such as gradient descent or genetic algorithms are employed to find suitable counterfactual instances. The generated counterfactuals address interpretability challenges [143].

6.4.2 Optimization-based methods

Another approach is the optimization-based method, which slightly differs from the perturbation-based method. After generating CF instances by solving the optimization problem (minimizing original and CF instance differences), the method focuses on satisfying certain constraints, such as individual and group fairness of the generated CFs. By applying this approach, Kusner et al. (2017) introduced a framework for generating counterfactual explanations by minimizing the distance in a latent feature space [127]. Besides them, Samali et al. developed an optimization technique to ensure fairness in methods by creating representations with similar richness for different groups in the dataset [144]. They represented experimental results showing that men's faces have lower reconstruction errors than women's in an image dataset. They developed a dimensionality reduction technique utilizing an optimization function mentioned in equation (2).

$$min_{U \in R^{m \times n}, rank(U) \leq d} max \left\{ \frac{1}{|A|} loss(A, U_A), \frac{1}{|B|} loss(B, U_B) \right\} \tag{2}$$

Here, A and B are two subgroups, U_A and U_B denote matrices whose rows correspond to rows of U, U contain members of subgroups A and B given m data points in R_n. Their proposed algorithm is summed up in two steps: firstly, it relaxes the objective to a semidefinite program (SDP) and solves it. Secondly, it solves a linear program that would reduce the rank of the solution [144].

6.4.3 Rule-based methods

Additionally, rule-based methods have been proposed, such as the Anchors algorithm by Ribeiro et al., which generates rule-based explanations by identifying the smallest set of features that must be true for a specific prediction [130].

6.4.4 Combining multiple methods

These different methods offer diverse ways to generate counterfactual explanations, allowing researchers and practitioners to choose the most suitable approach for their needs. Some other scholars emphasize generating diverse CFs to explore the explanation space and identify diverse and coherent explanations. This method combines multiple techniques such as heuristics, optimization algorithms, sampling methods for searching, and pruning techniques. It also captures the trade-off between diversity and coherence. It may penalize redundant or overlapping explanations while rewarding diverse and coherent explanations. Candidate explanations generated from this method are diverse and coherent [76].

6.4.5 Multi-modal alternative profiles

Besides the linear counterfactual generation methods mentioned above, scholars also explore multi-modal CF generation. For example, Abbasnejad et al. propose generating counterfactual instances by modifying both the input image and the generated text. These modifications capture alternative visual and linguistic explanations, resulting in different model predictions. This function typically includes terms encouraging visual fidelity, linguistic coherence, and dissimilarity from the original instance [43].

6.5 Fairness terminologies and metrics definitions:

The filtered articles proposed various fairness-related terminologies to mitigate fairness issues by implementing them in bias reduction strategies. We present generalized descriptions of these definitions.

- **Disparate impact:** Feldman et al. describe the disparate impact as a situation in which a decision-making process disproportionately impacts members of a protected group, regardless of intent [99] or in other words, disparate impact is a predictor that makes different errors for different feature groups [93, 125]. Disparate impact can be measured using statistical techniques such as the "disparate impact ratio", which compares the proportion of favorable outcomes (such as job offers) between different groups. If the ratio is significantly different between groups, it suggests that the model is exhibiting disparate impact. For example, from a dataset of three men and five women, this job offering algorithm offers jobs to two men and two women, then the ratio for men $\frac{2}{3}$ is significantly larger than $\frac{2}{5}$. It indicates the presence of disparate impact.
- **Causal fairness:** A decision rule is causally fair for a protected attribute if changing that protected attribute while holding all other variables constant does not change the probability of receiving a positive outcome [125]. In other words, the protected attribute should not cause any outcome differences. For example, in a dataset of people's information, such as age, gender, salary, and mortgage rate, if a credit card-allowing algorithm provides the exact prediction (allow credit card for that person) in the presence and absence of gender, then the algorithm has causal fairness.
- **Demographic parity:** Dwork et al. state that demographic parity is satisfied if the proportion of positive outcomes is equal across all groups [99]. Keeping the meaning same, Hardt et al. defined the demographic parity when the true positive rate (TPR) is equal across all groups [125]. Emphasizing the positive outcome, Feldman et al. claimed that a classifier has demographic parity if its positive predictive value (PPV) is equal across all groups [102]. For example, in job applications, if the model selects 10% of male and 10% of female candidates for interviews, then demographic parity is satisfied. Here, 10% is the TPR and PPV for both males and females, which is equal. The classifier predicts negative outputs for 90% males and females, indicating the negative predictive value.
- **Equalized odds and calibration:** In the context of fair machine learning, 'equalized odds' is a fairness criterion that requires the true positive and false positive rates to be equal across different groups defined by a sensitive attribute, such as race or gender [132]. True positive (TP) and false positive (FP) rates are commonly used performance metrics in classification tasks. Specifically, equalized odds means that the probability of a positive outcome (e.g., being approved for a loan or receiving a medical intervention) should be the same for individuals in different groups who have the same true status (e.g., whether they will pay back the loan or whether they have a particular medical condition).
- **Group fairness:** It is a popular fairness concept defined and explored by many researchers from several perspectives to implement it in developing fair model

[64, 97, 99, 125, 128]. In the definition of group fairness by Dwork et al., they emphasized the model to have demographic parity and not to have a disparate impact for ensuring that similar individuals be treated similarly, regardless of their group membership [99]. Similarly, Feldman defined group fairness regarding the disparate impact and stated that the ratio of positive outcomes (e.g., being released on parole) should be roughly equal across different groups defined by sensitive attributes (e.g., race or gender) [93]. Researchers also defined group fairness in terms of equalized odds. Such as Li et al. stated that group fairness requires that the true positive and false positive rates be the same across all groups [77]. It means the probability of a loan being approved should be the same for all groups, regardless of social relationships, race, gender, or other protected attributes [55, 77].

- **Individual fairness:** It refers that two individuals with similar relevant characteristics (e.g., credit history, income) should receive similar decisions [77]. To explain it with a previous example, from a dataset of three men and five women, a job offering algorithm offers jobs to two men with similar characteristics, such as a similar salary range and mortgage rate. However, if the third man has a similar salary and mortgage rate as the other two men, then we can expect that the model will also offer jobs to this third man if it has individual fairness.
- **Contrastive fairness:** It is a fairness criterion that focuses on comparing the outcomes of two groups that are similar in all relevant ways except for their membership in a protected attribute group, such as race or gender. The idea is to evaluate the extent to which their protected attribute status can explain the difference in outcomes between these groups and to ensure that this difference is not more significant than it would be if the groups were not distinguished by their protected attribute [42]. For example, in a hiring context, contrastive fairness would require that two equally qualified candidates with different protected attribute statuses (e.g., one male and one female) have roughly the same chance of being selected for the job. This fairness criterion addresses situations where the compared groups are not entirely distinct and non-protected attributes can not fully explain outcome differences.
- **Burden:** This terminology inclines to remove discrimination between categorical groups. It refers to the cost or effort required to achieve a particular level of transparency and fairness in a machine-learning model. There is often a trade-off between a model's level of interpretability and fairness and the cost of achieving these goals [73]. For example, adding interpretability or fairness constraints to a model may increase the computational cost of training or evaluating the model or require additional data collection or annotation, which refers to the burden. We must balance the burden of achieving a certain level of transparency or fairness against the potential benefits of using the model in a particular application.
- **Equality of effort:** This concept is analyzed as a causal-based fairness approach, which refers to a notion of fairness in which individuals are judged based on their effort and not just their outcomes. According to the context of group fairness in machine learning, equality of effort requires that individuals who put in the same amount of effort (most other features have the same value) should have equal chances of success, regardless of any protected group status (such as race or gender) [80].

For example, A female and a male candidate with similar qualifications should have equal chances of being approved for a job interview.

- **Causal probabilistic logic:** Scholars explore causal probabilistic logic to remove the ambiguity of the judgment-based causal fairness idea. Causal probabilistic logic is a type of logic that attempts to understand the relationships between cause and effect in probabilistic terms. One crucial aspect of causal probabilistic logic is the use of counterfactuals, which are statements about what would have happened if a particular event had not occurred. Counterfactuals allow one to reason about the causal effects of interventions and can be used to test causal hypotheses [50]. In this approach, causality is modeled probabilistically, where events are seen as causes of other events with some probability.
- **Group Fairness Indicator (GFI):** It refers to a metric that quantifies the level of fairness achieved for different groups or sub-populations within a given context. A GFI typically considers the outcomes or predictions made by the system for various groups and compares them based on a fairness criterion. The choice of fairness criterion varies depending on the specific context and the fairness principle being considered, such as equalized odds, demographic parity, or equal opportunity. For example, in the context of classification algorithms, a GFI compares the true positive rates, false positive rates, or predictive accuracy for different demographic groups (such as gender or race). By analyzing these metrics, researchers and practitioners can evaluate whether the algorithm exhibits disparities or biases in its predictions across different groups. The purpose of a group fairness indicator is to provide an objective measure of fairness and enable the identification of any unfairness or discriminatory patterns in the decision-making process. In addition, it helps stakeholders assess the performance of algorithms and make informed decisions to mitigate any observed disparities [45].
- **Proximity:** The definition of proximity is required for exploring CFs. It refers to the similarity between the original instance and the counterfactual instance generated to explain the prediction of a machine learning model. It measures how far the counterfactual instance is from the original instance in the feature space. For example, if a machine learning model predicts that a loan application will be denied, a counterfactual explanation could generate a new loan application similar to the original one but with some changes resulting in the loan being approved. The proximity of the counterfactual loan application would reflect how similar it is to the original loan application [70, 106].
- **Sparsity** Similar to proximity, sparsity, validity, and diversity are defined for CF explanations. Firstly, sparsity refers to the property of the explanation that suggests changing as few features as possible while still achieving the desired outcome. For example, a sparse example identifies a few key features of a loan application that change the outcome of the loan decision if we modify those key features [70].
- **Validity** Secondly, validity refers to the degree to which the counterfactual explanations provided for a given model's output are true and realistic. For example, suppose a CF for loan approval ML model's predictions say to increase the monthly income within two days for the approval to be granted. In that case, the CFs need more validity as it is impossible [70].

- **Diversity** Finally, Diversity measures how distinct the different counterfactual explanations are. In other words, diverse counterfactual explanations provide a range of alternative scenarios that explain the model's decision from multiple perspectives [70]. For example, suppose a machine learning model classifies a loan applicant as high risk due to their low credit score. A diverse set of counterfactual explanations could include providing alternative scenarios for the applicant to improve their credit score, such as paying off debt, getting a secured credit card, or taking out a credit-builder loan. By presenting various possible solutions, diverse explanations can help the user better understand the model's decision-making process and increase their trust in the system.

7 Challenges and limitations of the methodologies

Although these methodologies developed with fairness-related terminologies solve many issues, they also generate other challenges. Some scholars have addressed these drawbacks of their suggested approach, such as in which circumstances their method would only function sometimes. Some of the typical limitations of the proposed methodologies are the limited scope of protected attributes, the assumption of causal relationships, the failure to handle categorical features, the focus on individual fairness, ambiguous explanation, the data quality and representativeness, and the trade-off between fairness and accuracy. Fig. 10 represents these limitations and which methodologies can potentially have these limitations. The figure also represents an overall illustration of issues, the methods to solve them, and their limitations.

7.1 Limited scope of protected attributes

Many existing bias reduction methods focus on addressing bias related to a specific set of protected attributes, such as race or gender, while neglecting other potential sources of bias [64, 93, 98, 115, 124]. This limited scope may lead to insufficient mitigation of bias [97].

7.2 The assumption of causal relationships

Additionally, causal relationship assumptions can generate unwanted bias in the model prediction. Experts generally researched counterfactual explanations in light of two mechanisms: contrast effects and causal inferences. However, these suggested methods are temporary and costly [54]. Moreover, Causal inference models offer to answer causal questions like, "If a feature A changes, what will happen to outcome y" [60]. However, causal inference does not describe how the confidence of a particular outcome is altered for ignorable changes in the training set [84]. The difficulties in generating counterfactuals may also lead to undesired fairness challenges [54].

7.3 The failure to handle categorical features

Also, developing methodologies by exploring similar datasets reduces the chance of handling different features. Many existing methodologies work with datasets that mostly have continuous features. Thus, Existing CF generating algorithms may fail

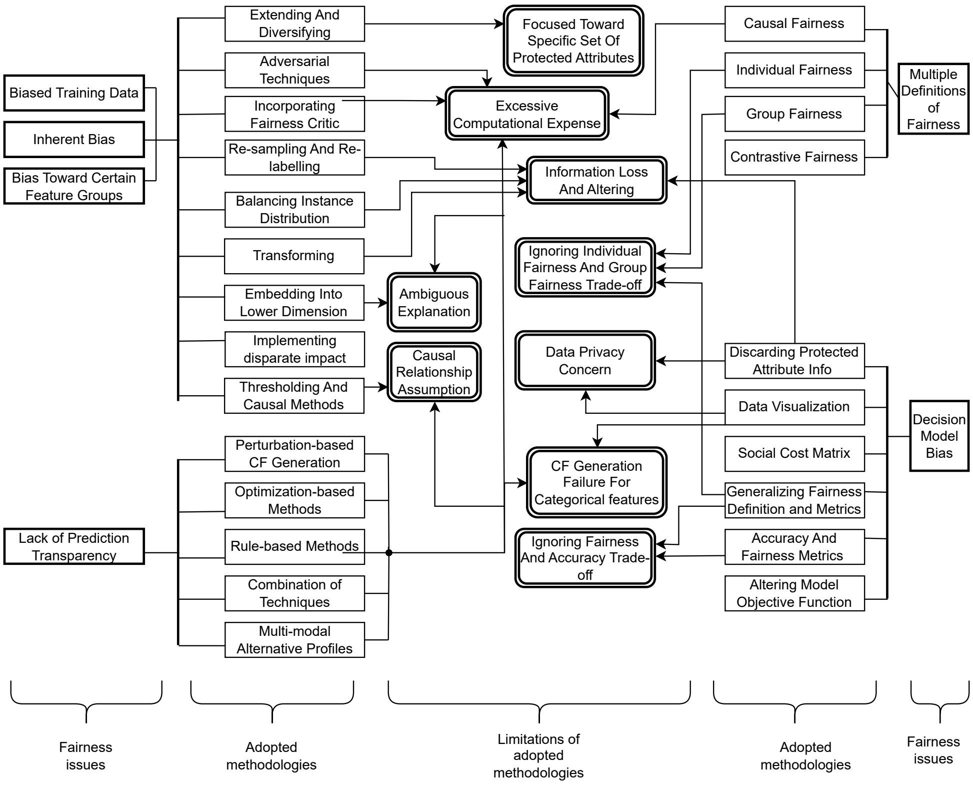
Biased Training Data
Inherent Bias
Bias Toward Certain Feature Groups
Extending And Diversifying
Adversarial Techniques
Incorporating Fairness Critic
Re-sampling And Re-labelling
Balancing Instance Distribution
Transforming
Embedding Into Lower Dimension
Implementing disparate impact
Thresholding And Causal Methods
Focused Toward Specific Set Of Protected Attributes
Excessive Computational Expense
Information Loss And Altering
Ambiguous Explanation
Causal Relationship Assumption
Ignoring Individual Fairness And Group Fairness Trade-off
Data Privacy Concern
CF Generation Failure For Categorical features
Ignoring Fairness And Accuracy Trade-off
Lack of Prediction Transparency
Perturbation-based CF Generation
Optimization-based Methods
Rule-based Methods
Combination of Techniques
Multi-modal Alternative Profiles
Causal Fairness
Individual Fairness
Group Fairness
Contrastive Fairness
Multiple Definitions of Fairness
Discarding Protected Attribute Info
Data Visualization
Social Cost Matrix
Generalizing Fairness Definition and Metrics
Accuracy And Fairness Metrics
Altering Model Objective Function
Decision Model Bias
Fairness issues
Adopted methodologies
Limitations of adopted methodologies
Adopted methodologies
Fairness issues

Fig. 10 An overview of the classification of adopted methodologies targeted to solve several issues and their limitations (from left to middle and right to middle).

to handle categorical features [72]. This restriction to categorical or group features results in the misled measurement of fairness [77].

7.4 The focus on individual fairness

Next, several bias reduction approaches prioritize individual fairness, which aims to treat similar individuals similarly. However, they may overlook group fairness, resulting in disparate outcomes for certain groups [145].

7.5 Ambiguous explanation

Furthermore, CFs may occasionally fail to ensure logical explanations. In reality, a set of diverse counterfactuals may contain changes of attribute values for altering the prediction that is not changeable to those values [69]. In this aspect, research is scarce regarding what should be optimized to generate a feasible set of counterfactual [69, 70].

7.6 The data privacy concern

Along with unfeasible CF explanations, CF-generating approaches trigger provoking protected attributes [69]. Discrimination detection algorithms are developed based on the assumption that all attribute information is visible to algorithms. This assumption is not valid for all scenarios. Thus, these algorithms are limited to getting protected features [55, 77, 80].

7.7 The trade-off between fairness and accuracy

Besides these limitations, data quality also contributes to challenges. For example, bias reduction methods rely heavily on the training data's quality and representativeness. If the training data contains biases or is unrepresentative of the target population, the methods may not effectively mitigate bias [95]. Finally, Some bias reduction techniques involve modifying the model or training process to achieve fairness, which can lead to a trade-off between fairness and accuracy. Striving for perfect fairness may come at the cost of decreased predictive performance [126].

7.8 The information loss and altering

The goal of re-sampling is to ensure that we train the model on a balanced dataset, which can help mitigate bias and improve the model's overall performance. However, it is essential to note that re-sampling can also result in a loss of information, and we need to ensure that the re-sampled dataset is representative of the original dataset.

7.9 The excessive computational expense

Many of the adopted methodologies involve adversarial techniques, and the main problem with adversarial techniques is that they can be computationally expensive. Additionally, and may not always be effective in addressing all forms of bias. Adversarial training, or adversarial debiasing, involves generating adversarial examples in various ways, such as perturbing the input data to the model to maximize the model's loss or minimize its accuracy. This process requires solving an optimization problem for each training example, which can be computationally expensive, especially for large datasets. Moreover, adversarial techniques may require additional training iterations to achieve convergence, leading to a further computational burden. Additionally, generating adversarial examples may require running the model multiple times for each example, increasing the computational cost. Finally, adversarial techniques may require specialized hardware or software to efficiently generate adversarial examples, adding to the computational expense.

These limitations highlight the challenges, and we need to adopt and implement methodologies to address these limitations in developing and applying bias reduction methodologies.

8 Future Direction

Some of our studied papers proposed future research directions (mainly in the conclusion or introduction of the paper). We gathered all this information while doing our mapping analysis and classified these indicated future efforts into a few categories, which we will explain next.

8.1 Fairness methodologies for models beyond binary decisions:

Most filtered research articles proposed methodologies to ensure fairness for models that only perform binary prediction [98, 125, 132]. For example, a credit card denying/accepting model predicts only 'yes' indicating credit card request accepted and 'no' indicating credit card request rejected [125, 146]. However, expanding fairness beyond binary decisions is a future direction. It includes addressing fairness in multi-class classification and regressive tasks. For example, predicting the right insurance plan, such as 'start-up family pack', 'small family pack', or 'large family pack' for a family, based on the earning member's income, requires a model with multi-class classification. Also, we may need regressive models to settle an amount for offering salary for an individual depending on his/her qualification and company requirement, which also requires fairness for all candidates. In the case of a lower initial offering, many competitive candidates may not even feel the need to negotiate based on the offering. In contrast, in the case of a high initial offering, the company may suffer in the long run with lower potential or lower employee performance. To ensure fairness, the regressive model should have minor differences in initial salary offerings for candidates with the same qualifications but different age ranges, races, or genders. Thus, developing methodologies that account for nuanced differences among groups rather than focusing solely on binary outcomes can be a noteworthy contribution in this field [147].

8.2 Intersectionality-Aware Fairness methodologies for protected features:

Most of the fairness-ensuring strategies explained in the filtered papers focused on reducing bias towards a protected feature in the dataset [63, 64, 93, 115, 124, 124]. Thus, traditional fairness ensuring strategies often focus on individual protected attributes in isolation, assuming that we can address the biases associated with each attribute separately. However, intersectionality-aware fairness methodologies address biases that emerge from the intersections of multiple attributes. It recognizes multi-dimensional and interconnected social identities and discriminations arising from the combination of multiple protected attributes, such as race, gender, age, and socioeconomic status. For example, consider a credit card application process where decisions are made based on various attributes, such as income, employment status, and credit history. Intersectionality-Aware Fairness would consider how biases and discrimination may arise when individuals possess multiple intersecting attributes, such as a woman of color or a low-income transgender individual. They may face unique forms of bias that cannot be adequately captured by considering each attribute in isolation. Thus, by considering intersectionality awareness, RAI tools and other bias reduction approaches can better capture the multidimensional and interconnected nature of social identities and address the biases and discrimination that arise from the combination of multiple protected attributes [87].

8.3 User-Defined Fairness and Customization:

Existing fairness-ensuring methodologies maintain two steps: 1. defining fairness from social, statistical, or other perspectives and 2. ensuring the defined fairness in the proposed method. Researchers or policymakers pre-define fairness definitions, imposing a one-size-fits-all notion of fairness on algorithmic decision-making. However, different individuals and communities may have different perspectives, values, and priorities regarding fairness. Allowing users to customize fairness definitions and constraints can provide a more inclusive and personalized approach to fairness. User-defined fairness and customization involve empowering individuals to have a say in defining their notions of fairness and incorporating their preferences into the fairness-ensuring methodologies. We can tailor methodologies to align with individual perspectives and values by enabling users to define their fairness criteria. This customization can take various forms, such as providing adjustable fairness thresholds, allowing users to prioritize different groups or attributes, or incorporating feedback mechanisms to refine fairness definitions based on user input iteratively. The idea of user-defined fairness and customization reflects the importance of fairness being context-dependent and subjective to some extent. It acknowledges that fairness is a complex and multidimensional concept that should be adaptable to different peoples' and communities' specific needs and preferences.

8.4 fairness ensuring methodologies considering long-term concept definitions and dynamics:

The future direction also involves expanding fairness-ensuring methodologies to consider the effects of interventions and algorithmic decisions over time. This direction recognizes that fairness is not a static concept and that disparities may emerge or change in different contexts and timeframes. Methodologies must examine how interventions and algorithmic decisions impact fairness outcomes over extended periods to address long-term fairness. It requires understanding the dynamics of fairness and considering how biases and disparities can manifest or evolve. Additionally, long-term fairness involves accounting for the potential unintended consequences of interventions and algorithmic systems. Fairness-ensuring methodologies should assess the long-term effects of such interventions to ensure that they do not inadvertently reinforce or introduce new biases or disparities [148].

8.5 Preparing and publishing unbiased datasets:

Researchers must focus on removing bias in popular datasets to promote fairness in the models developed from these datasets. Many researchers search for a dataset free of intricate biases as the data and the state of the dataset's attribute can be biased [54, 55, 70]. They need these datasets for testing the fairness of RAIs or other predicting models. Scholars have introduced approaches to test if a model prediction is biased toward any group [110, 111]. However, if we apply these approaches to predictive models with biased datasets, the results may not indicate that even though the model is fair. This situation will make the purpose of the unfairness testing algorithms ambiguous. Thus, if some scholars remove some biases from a few datasets and make them publicly available, other scholars can look into them and work on removing other biases from those datasets. These datasets can be widely explored for developing models without worrying about unfair models.

9 Source Code, Datasets, Tools

Some filtered studies have developed tools to contribute to model fairness research and represent the results of implementing their approach in standard datasets to prove their claim. We considered the accessibility of these datasets and special proposed tools if they have provided a source code repository. Some researchers also pointed to dataset repositories that are not publically accessible. We present these tools and the popular datasets these articles explore in section 9.1 and 9.

9.1 Developed Tools/Frameworks

- *Aequitas* [129]: This toolkit generates reports from the obtained data to test if an ML model is fair for different subgroups. Aequitas can help people from various professions, such as data scientists, ML researchers, and policymakers.
- *AIF360*[1] [149]: AI Fairness 360 or AIF360 is an industrial Python toolkit developed by IBM mainly for evaluating fairness algorithms and providing a common framework so that scholars can share their ideas. A complete collection of fairness metrics for datasets and models, justifications for these metrics, and dataset's and model's bias mitigation strategies are included in the package along with an interactive Web experience.
- *DiCE*[2] [70]: It is an open-source quantitative evaluation framework for counterfactuals that allows fine-tuning for a particular scenario and enables comparison between CF-based and other local explanation based methods [150]. This tool provides diverse counterfactual instances that are different from the original but represent the same class. This method used diversity metrics and proximity constraints for generating diverse and feasible CFs (where the prediction is binary).
- *ViCE*[3]: It is a black-box visual analytic tool that enhances the interpretability of machine learning models in the context of visual tasks. The approach focuses on generating counterfactual explanations by providing alternative visual examples that would lead to different model predictions. It uses a heuristic search algorithm, a Gaussian technique in features, and a greedy approach to discover the lowest set of viable adjustments for changing the outcome [72].

[1] https://github.com/Trusted-AI/AIF360
[2] https://github.com/microsoft/DiCE
[3] https://github.com/5teffen/ViCE

- *CERTIFAI*: The development of counterfactual explanations is the main focus of CERTIFAI (Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence models) [73]. Existing counterfactual generating algorithms have a few shortcomings, such as infeasible examples. It hinders the process of evaluating the robustness and fairness of developed models. This tool aims to provide a solution to this issue.

9.2 Supporting Datasets

- *HELOC* [151]: The HELOC (Home Equity Line of Credit) dataset is a real-world dataset popularly known in the field of credit risk assessment and lending. The HELOC (Home Equity Line of Credit) dataset is a real-world dataset commonly used in credit risk assessment and lending. It contains demographic information about borrowers' credit profiles, loan applications, and loan performance related to home equity lines of credit. Researchers often utilize this dataset for developing and evaluating machine learning models to predict credit risk, determine the likelihood of defaulting on a loan, improve lending decisions, manage credit risk, and assess the fairness of loan approval processes. For example, Gomez et al. ran their case study to show how their proposed technique provided the ML explanation for prediction on this dataset [72].
- *Child welfare dataset*: This dataset contains child welfare and protection information. The specific attributes and labels in the dataset may vary depending on the source and purpose of the dataset. However, there are usually demographic attributes (age, gender, ethnicity), socioeconomic attributes (income, education level), case information (referral, abuse or neglect reported), family history (parental substance abuse, domestic violence, or mental health issues), placement history (foster placement history), service utilization (counseling, therapy, or parenting programs). In the academic field, the Child Welfare dataset is a valuable resource for understanding the ethical dimensions of child welfare practices and the implications of using AI technologies in this domain. It allows researchers to identify potential challenges, propose solutions, and contribute to developing ethical guidelines and frameworks for AI systems in child welfare. For example, Mishler showed Risk Assessment Instruments tools' inability to generate separate risk scores with the help of this dataset [67].
- *UCI adult dataset* [152]: The UCI Adult dataset, also known as the Census Income dataset, is a popular dataset used in machine learning and data mining research. The UCI Adult dataset consists of 14 attributes or features. These attributes capture various demographic, social, and economic information about individuals. The type of attributes includes a mix of categorical and numerical attributes. Categorical attributes represent specific characteristics such as education, marital status, occupation, and relationship status. Numerical attributes include age, educational years, capital gain, loss, and weekly work hours. The predicting attribute in the UCI Adult dataset is typically the "income" attribute, which indicates whether an individual earns more than $50,000$ per year. This attribute serves as a binary label, often used for classification tasks. The UCI Adult dataset is valued for its real-world

relevance, the presence of socio-economic attributes, and the opportunities it provides for studying fairness, bias, and income prediction tasks. Its availability and well-documented nature make it a suitable dataset for many field researchers. For example, Ramaravind K. Mothilal et al. evaluated their method of explaining the ML model on this dataset [70]. Sharma et al. used this dataset to generate and analyze a comparison of the same person's explanation to demonstrate their model's superiority [73]. Furthermore, another paper used seven attributes (sex as the protected attribute, age, marital status, work class, education, hours, and income as outcome) of this dataset to evaluate their proposed discrimination detection and removal algorithms based on equality of effort [80].

- *UCI German Credit Dataset* [153] and *Dataset from Lending Club* [154]: The UCI German Credit Dataset is a well-known dataset used in machine learning and credit risk analysis. It consists of 20 attributes or features (a mix of categorical and numerical attributes) that capture various aspects of individuals applying for credit, including personal, financial, and employment information. Categorical attributes include sex, housing status, employment type, and credit history. Numerical attributes include features like age, credit amount, duration of credit, and installment rate. The predicting attribute in the UCI German Credit Dataset is typically the "credit risk" attribute, which indicates whether a person is considered a good or bad credit risk. This binary label is used for classification tasks to predict the creditworthiness of applicants. Similar to the UCI German credit dataset, the dataset from Lending Club typically consists of several attributes (categorical and numerical). It includes numerous attributes that provide information about loan applicants and their financial profiles. The predicting attribute is typically the loan status or loan outcome. This attribute indicates whether a loan was fully paid, charged off, in default, or had another status. It can be used for classification tasks to predict the likelihood of loan default or assess lending models' performance. To evaluate the model accuracy and feature information of loan decisions with the proposed model and other models, these datasets were used by [70].
- *Diabetes Dataset* [155]: It is widely used in healthcare research. the key characteristics of this dataset are nine attributes capturing health-related measurements and demographic information (numerical and categorical). The predicting attribute in the Diabetes Dataset is the "Outcome" attribute, which represents the presence or absence of diabetes in the individual. This attribute serves as the target variable for classification tasks. Gomez [72] et al. explained an instance in this database and contextualized the values of the dataset.
- *Pima Indian diabetes dataset*: For comparing the robustness of different models, S. Sharma [73] et al. utilized this dataset. The dataset serves as a benchmark for evaluating the performance of different classification algorithms in the context of diabetes prediction. Researchers can compare their models' accuracy, sensitivity, specificity, and other metrics with existing literature that utilizes this dataset. The simplicity and interpretability of the Diabetes Dataset also make it suitable for students and beginners to practice and understand the concepts of data preprocessing, feature selection, model training, and evaluation in a healthcare context.

- *Outbrain Click Prediction* [156] and *KKBox's Music Recommendation Challenge* [157]: These datasets are available in Kaggle. To compare test scores using different positional approaches, Yuan et al. used these datasets [158].
- *HMDA* [4]: The Home Mortgage Disclosure Act (HMDA) dataset is a collection of data related to mortgage applications and loans in the United States. It contains information on various attributes related to loan applications, borrowers, lenders, and loan characteristics. The dataset provides valuable insights into lending practices and can be used to analyze mortgage market trends, identify potential disparities or biases, and assess fair lending practices. The attributes included in the HMDA dataset can vary depending on the year and jurisdiction. However, typical attributes in HMDA datasets include applicant's information, load information, and lender information. The dataset also includes information on the loan approval status, denials, and other loan-related outcomes, which can be used as labels for predicting loan outcomes or assessing fairness [56, 159]. The official website of the Consumer Financial Protection Bureau (CFPB), the organization responsible for collecting and maintaining the HMDA data, can be an excellent source for accessing the dataset through their public data platform.

10 Threats to validity of our study

We attempted to include papers on the fairness study of machine learning prediction using counterfactual notions with our query. However, owing to the limits of our query, there is still a chance that we may miss out on considerable research. We used the same query for all repositories, but the terms' scope differed in a few cases. For example, for ACM DL, we used "machine learning" and "fairness" within the whole article and "counterfactual" within only the abstract of the article. However, for IEEE Xplore, this threshold of query terms resulted in 2000+ search results that could not be processed. As a result, we set the query words boundary to be within the article's abstract. We also investigated related works of these publications to filter out as many significant articles as feasible. However, it is still easy to overlook certain essential linked studies.

Furthermore, we've concentrated on skimming through articles to address the research questions mentioned in Section 3. Our review strategy offers an overall view of the field by outlining categories of fairness issues, adopted methodologies, and their limitations. By diving deeper into the different methodologies used, we can engage in a detailed discussion on potential developments in ensuring fairness. While the proposed taxonomy and current discussions in our paper suit a review aimed at introducing newcomers to the field, categorizing methodologies from various perspectives, such as based on implementation time perspectives—like in-processing, pre-processing, and post-processing [30]—provides intriguing insights for researchers focusing on method development.

[4] https://www.consumerfinance.gov/data-research/hmda/historic-data/

11 Conclusion

We followed a systematic approach to explore the current research trends in this field. First, we examined the present method of performing systematic mapping studies followed by other scholars. Then, we proceeded to the section where we followed the rules for portraying secondary research work in a classified and informative structure. The majority of the procedures we are following in this study are based on best practices outlined by Wieringa *et al.* [34], Das *et al.* [35], Petersen *et al.* [160], Kitchenham *et al.* [33], Gonçales *et al.* [161]. Following the steps and best literature review practices, we can summarize our contributions as follows:

- We classified and synthesized articles depending on generalizing the types of approaches scholars explored.
- We constructed the research questions for our study and structured our generalized query for four popular databases.
- We explicitly discussed our search results with the query in the form of our research question-answer and referenced relevant articles that the keyword search did not include.

From the study, we conclude that a model with high accuracy can represent multiple types of fairness issues, such as bias against protected attributes, inherent data bias, or lack of explanation. Different types of bias require different types of approaches. Bias mitigation methodologies are only partially immune to these biases. Handling numerous fairness issues in one model may result in a new and distinctive fairness issue [84]. As a result, understanding the current need to ensure model fairness requires a thorough study of the previous methods and their difficulties. Thus, generalizing the fairness issues and classifying the methodologies from the perspective of these issues may contribute to improving the existing methodologies and developing advanced methodologies. So, we contributed in this regard and summarized our contribution as follows.

- We provide insights into the current landscape of fairness by highlighting the issues scholars are exploring. We generalize the issues into six groups and discuss the key factors contributing to these issues.
- We classify and discuss the adopted methodologies to solve these issues highlighting how we can mitigate training data bias, mitigate bias toward protected attributes, and provide prediction explanation and interpretation.
- Furthermore, we organize the challenges of these methods and link the discussed challenges to these methodologies.
- We also contributed by discovering possible future directions from these articles.

In addition to the highlighted contributions, our mapping study methodology holds significant potential for future perspectives. While the article reviews offer insightful guidance for newcomers in the field, the systematic mapping approach detailed in Section 3 streamlines the process for researchers to review the current literature landscape (such as query development, searching databases with the query, filtering articles, etc.). Given that the articles surveyed in this paper might become outdated due to emerging methodologies over time. However, the mapping study approach will provide a reliable direction for guiding the review of newer methods in this domain.

Acknowledgments. The authors want to thank the reviewers for their thoughtful feedback and suggestions during the review process.

Declarations

Funding. Part of this work was performed while P.R. was funded by the National Science Foundation under Grant Nos. 2039678, 2136961, and 2210091. The views expressed herein are solely those of the author(s) and do not necessarily reflect those of the National Science Foundation.

Conflict of interest. The authors have no relevant financial or non-financial interests to disclose.

Data availability. Data sharing does not apply to this article as no datasets were generated or analyzed during the current study. This article is a systematic literature review, and as such, it synthesizes and analyzes information from previously published literature.

References

[1] Waters, A., Miikkulainen, R.: Grade: Machine-learning support for graduate admissions. AI Magazine **35**(1), 64–75 (2014) https://doi.org/10.1609/aimag.v35i1.2504

[2] Pessach, D., Singer, G., Avrahami, D., Chalutz Ben-Gal, H., Shmueli, E., Ben-Gal, I.: Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming. Decision Support Systems **134**, 113290 (2020) https://doi.org/10.1016/j.dss.2020.113290 . Accessed 2021-07-06

[3] Berkelaar, B.L., Buzzanell, P.M.: Online Employment Screening and Digital Career Capital: Exploring Employers' Use of Online Information for Personnel Selection. Management Communication Quarterly **29**(1), 84–113 (2015) https://doi.org/10.1177/0893318914554657 . Accessed 2021-07-06

[4] Jeske, D., Shultz, K.S.: Using social media content for screening in recruitment and selection: pros and cons. Work, Employment and Society **30**(3), 535–546 (2016) https://doi.org/10.1177/0950017015613746 . Accessed 2021-07-06

[5] Andini, M., Ciani, E., Blasio, G.d., D'Ignazio, A., Salvestrini, V.: Targeting policy-compliers with machine learning: an application to a tax rebate programme in Italy. Technical Report 1158, Bank of Italy, Economic Research and International Relations Area (December 2017). https://ideas.repec.org/p/bdi/wptemi/td_1158_17.html Accessed 2021-07-06

[6] Athey, S.: Beyond prediction: Using big data for policy problems. Science **355**(6324), 483–485 (2017) https://doi.org/10.1126/science.aal4321 . Accessed 2021-07-06

[7] Dai, W., Brisimi, T.S., Adams, W.G., Mela, T., Saligrama, V., Paschalidis, I.C.: Prediction of hospitalization due to heart diseases by supervised learning methods. International journal of medical informatics **84**(3), 189–197 (2015) https://doi.org/10.1016/j.ijmedinf.2014.10.002 . Accessed 2021-07-06

[8] Zliobaite, I.: A survey on measuring indirect discrimination in machine learning. arXiv preprint arXiv:1511.00148 (2015) https://doi.org/10.48550/arXiv.1511.00148

[9] Veale, M., Binns, R.: Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society **4**(2), 2053951717743530 (2017) https://doi.org/10.1177/2053951717743530

[10] Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y.: AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. arXiv (2018) https://doi.org/10.48550/arXiv.1810.01943

[11] Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data **5 2**, 153–163 (2016) https://doi.org/10.1089/big.2016.0047

[12] Corbett-Davies, S., Goel, S.: The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. arXiv (2018) https://doi.org/10.48550/arXiv.1808.00023

[13] Verma, S., Dickerson, J., Hines, K.: Counterfactual Explanations for Machine Learning: A Review. arXiv (2020) https://doi.org/10.48550/arXiv.2010.10596

[14] Zhang, Q., Zhang, X., Liu, Y., Wang, H., Gao, M., Zhang, J., Guo, R.: Debiasing recommendation by learning identifiable latent confounders. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. KDD '23, pp. 3353–3363. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3580305.3599296

[15] DENG, Y., Luo, P., Loy, C.C., Tang, X.: Pedestrian attribute recognition at far distance. In: Proceedings of the 22nd ACM International Conference on Multimedia. MM '14, pp. 789–792. Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2647868.2654966

[16] Choraś, M., Pawlicki, M., Puchalski, D., Kozik, R.: Machine learning – the

results are not the only thing that matters! what about security, explainability and fairness? In: Krzhizhanovskaya, V.V., Závodszky, G., Lees, M.H., Dongarra, J.J., Sloot, P.M.A., Brissos, S., Teixeira, J. (eds.) Computational Science – ICCS 2020, pp. 615–628. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-50423-6_46

[17] Jui, T.D., Bejarano, G.M., Rivas, P.: A machine learning-based segmentation approach for measuring similarity between sign languages. In: Efthimiou, E., Fotinea, S.-E., Hanke, T., Hochgesang, J.A., Kristoffersen, J., Mesch, J., Schulder, M. (eds.) Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources, pp. 94–101. European Language Resources Association (ELRA), Marseille, France (2022). https://www.sign-lang.uni-hamburg.de/lrec/pub/22018.pdf

[18] Adeyanju, I.A., Bello, O.O., Adegboye, M.A.: Machine learning methods for sign language recognition: A critical review and analysis. Intelligent Systems with Applications **12**, 200056 (2021) https://doi.org/10.1016/j.iswa.2021.200056

[19] Biswas, D., Tešić, J.: Small object difficulty (sod) modeling for objects detection in satellite images. In: 2022 14th International Conference on Computational Intelligence and Communication Networks (CICN), pp. 125–130 (2022). https://doi.org/10.1109/CICN56167.2022.10008383

[20] Biswas, D., Tevsi'c, J.: Progressive domain adaptation with contrastive learning for object detection in the satellite imagery. (2022). https://api.semanticscholar.org/CorpusID:255941648

[21] Knell, R.: On the analysis of non-linear allometries. Ecological Entomology **34**, 1–11 (2009) https://doi.org/10.1111/j.1365-2311.2008.01022.x

[22] Jui, T., Ayoade, O., Rivas, P., Orduz, J.: Performance analysis of quantum machine learning classifiers. In: NeurIPS 2021 Workshop LatinX in AI (2021). https://openreview.net/forum?id=oMEQXfmKshr

[23] Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016) https://doi.org/10.48550/arXiv.1609.02907

[24] Rivas, P., Thompson, C., Tafur, B., Khanal, B., Ayoade, O., Jui, T.D., Sooksatra, K., Orduz, J., Bejarano, G.: Chapter 15 - ai ethics for earth sciences. In: Sun, Z., Cristea, N., Rivas, P. (eds.) Artificial Intelligence in Earth Science, pp. 379–396. Elsevier, PA, USA (2023). https://doi.org/10.1016/B978-0-323-91737-7.00007-4

[25] Iosifidis, V., Fetahu, B., Ntoutsi, E.: Fae: A fairness-aware ensemble framework, pp. 1375–1380 (2019). https://api.semanticscholar.org/CorpusID:211011092

[26] Kamiran, F., Calders, T.: Classifying without discriminating. In: 2009 2nd International Conference on Computer, Control and Communication, pp. 1–6 (2009). https://doi.org/10.1109/IC4.2009.4909197

[27] Menon, A.K., Williamson, R.C.: The cost of fairness in classification. arXiv preprint arXiv:1705.09055 (2017) https://doi.org/10.48550/arXiv.1705.09055

[28] Adel, T., Valera, I., Ghahramani, Z., Weller, A.: One-network adversarial fairness. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 2412–2420 (2019). https://doi.org/10.1609/aaai.v33i01.33012412

[29] Binns, R.: Fairness in machine learning: Lessons from political philosophy. In: Conference on Fairness, Accountability and Transparency, pp. 149–159 (2018). PMLR. https://proceedings.mlr.press/v81/binns18a.html

[30] Caton, S., Haas, C.: Fairness in machine learning: A survey. ACM Comput. Surv. (2023) https://doi.org/10.1145/3616865

[31] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**(6) (2021) https://doi.org/10.1145/3457607

[32] Chen, Z., Zhang, J.M., Hort, M., Sarro, F., Harman, M.: Fairness testing: A comprehensive survey and analysis of trends. arXiv preprint arXiv:2207.10223 (2022) https://doi.org/10.48550/arXiv.2207.10223

[33] Kitchenham, B., Brereton, P.: A systematic review of systematic review process research in software engineering. Information and Software Technology **55**(12), 2049–2075 (2013) https://doi.org/10.1016/j.infsof.2013.07.010

[34] Wieringa, M.: What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20, pp. 1–18. Association for Computing Machinery, Barcelona, Spain (2020). https://doi.org/10.1145/3351095.3372833

[35] Das, D., Schiewe, M., Brighton, E., Fuller, M., Cerny, T., Bures, M., Frajtak, K., Shin, D., Tisnovsky, P.: Failure Prediction by Utilizing Log Analysis: A Systematic Mapping Study. In: Proceedings of the International Conference on Research in Adaptive and Convergent Systems. RACS '20, pp. 188–195. Association for Computing Machinery, Gwangju, Republic of Korea (2020). https://doi.org/10.1145/3400286.3418263

[36] Creswell, J.W., Creswell, J.D.: Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Sage publications, CA, USA (2017). https://www.amazon.com/Research-Design-Qualitative-Quantitative-Approaches/dp/1452226105

[37] Booth, W.C., Colomb, G.G., Williams, J.M.: The Craft of Research. University of Chicago press, IL, USA (2003). https://www.amazon.com/Research-Chicago-Writing-Editing-Publishing/dp/022623973X

[38] Petersen, K., Vakkalanka, S., Kuzniarz, L.: Guidelines for conducting systematic mapping studies in software engineering: An update. Information and Software Technology **64**, 1–18 (2015) https://doi.org/10.1016/j.infsof.2015.03.007

[39] Wohlin, C., Runeson, P., da Mota Silveira Neto, P.A., Engström, E., do Carmo Machado, I., de Almeida, E.S.: On the reliability of mapping studies in software engineering. Journal of Systems and Software **86**(10), 2594–2610 (2013) https://doi.org/10.1016/j.jss.2013.04.076

[40] Zhang, W., Zhang, M., Zhang, J., Liu, Z., Chen, Z., Wang, J., Raff, E., Messina, E.: Flexible and adaptive fairness-aware learning in non-stationary data streams. In: 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI), pp. 399–406 (2020). https://doi.org/10.1109/ICTAI50040.2020.00069

[41] Altman, M., Wood, A., Vayena, E.: A harm-reduction framework for algorithmic fairness. IEEE Security Privacy **16**(3), 34–45 (2018) https://doi.org/10.1109/MSP.2018.2701149

[42] Chakraborti, T., Patra, A., Noble, J.A.: Contrastive fairness in machine learning. IEEE Letters of the Computer Society **3**(2), 38–41 (2020) https://doi.org/10.1109/LOCS.2020.3007845

[43] Abbasnejad, E., Teney, D., Parvaneh, A., Shi, J., Hengel, A.: Counterfactual vision and language learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10041–10051 (2020). https://doi.org/10.1109/CVPR42600.2020.01006

[44] Kamiran, F., Karim, A., Zhang, X.: Decision theory for discrimination-aware classification. In: 2012 IEEE 12th International Conference on Data Mining, pp. 924–929 (2012). https://doi.org/10.1109/ICDM.2012.45

[45] Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR 2011, pp. 1521–1528 (2011). https://doi.org/10.1109/CVPR.2011.5995347

[46] Kairouz, P., Liao, J., Huang, C., Vyas, M., Welfert, M., Sankar, L.: Generating Fair Universal Representations Using Adversarial Models (2022). https://doi.org/10.1109/TIFS.2022.3170265

[47] Calmon, F.d.P., Wei, D., Vinzamuri, B., Ramamurthy, K.N., Varshney, K.R.: Data pre-processing for discrimination prevention: Information-theoretic optimization and analysis. IEEE Journal of Selected Topics in Signal Processing **12**(5), 1106–1119 (2018) https://doi.org/10.1109/JSTSP.2018.2865887

[48] Kim, B., Park, J., Suh, J.: Transparency and accountability in ai decision support: Explaining and visualizing convolutional neural networks for text information. Decision Support Systems **134**, 113302 (2020) https://doi.org/10.1016/j.dss.2020.113302

[49] Riveiro, M., Thill, S.: "that's (not) the output i expected!" on the role of end user expectations in creating explanations of ai systems. Artificial Intelligence **298**, 103507 (2021) https://doi.org/10.1016/j.artint.2021.103507

[50] Beckers, S., Vennekens, J.: A general framework for defining and extending actual causation using cp-logic. International Journal of Approximate Reasoning **77**, 105–126 (2016) https://doi.org/10.1016/j.ijar.2016.05.008

[51] Nicklin, J.M., Greenbaum, R., McNall, L.A., Folger, R., Williams, K.J.: The importance of contextual variables when judging fairness: An examination of counterfactual thoughts and fairness theory. Organizational Behavior and Human Decision Processes **114**(2), 127–141 (2011) https://doi.org/10.1016/j.obhdp.2010.10.007

[52] Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019) https://doi.org/10.1016/j.artint.2018.07.007

[53] Ganegoda, D.B., Folger, R.: Framing effects in justice perceptions: Prospect theory and counterfactuals. Organizational Behavior and Human Decision Processes **126**, 27–36 (2015) https://doi.org/10.1016/j.obhdp.2014.10.002

[54] Roese, N.: Counterfactual thinking and decision making. Psychonomic Bulletin & Review **6**(4), 570–578 (1999) https://doi.org/10.3758/BF03212965 . Accessed 2021-07-17

[55] Balayn, A., Lofi, C., Houben, G.-J.: 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) https://doi.org/10.1007/s00778-021-00671-8 . Accessed 2021-07-17

[56] Lee, M.S.A., Floridi, L.: Algorithmic Fairness in Mortgage Lending: from Absolute Conditions to Relational Trade-offs. Minds and Machines **31**(1), 165–191 (2021) https://doi.org/10.1007/s11023-020-09529-4 . Accessed 2021-07-17

[57] Gulshad, S., Smeulders, A.: Counterfactual attribute-based visual explanations for classification. International Journal of Multimedia Information Retrieval **10**(2), 127–140 (2021) https://doi.org/10.1007/s13735-021-00208-3 . Accessed 2021-07-17

[58] Mellem, M.S., Kollada, M., Tiller, J., Lauritzen, T.: Explainable AI enables

clinical trial patient selection to retrospectively improve treatment effects in schizophrenia. BMC Medical Informatics and Decision Making **21**(1), 162 (2021) https://doi.org/10.1186/s12911-021-01510-0 . Accessed 2021-07-17

[59] Watson, D.S., Floridi, L.: The explanation game: a formal framework for interpretable machine learning. Synthese (2020) https://doi.org/10.1007/s11229-020-02629-9 . Accessed 2021-07-17

[60] Bertoncello, A., Oppenheim, G., Cordier, P., Gourvénec, S., Mathieu, J.-P., Chaput, E., Kurth, T.: Using Causal Inference in Field Development Optimization: Application to Unconventional Plays. Mathematical Geosciences **52**(5), 619–635 (2020) https://doi.org/10.1007/s11004-019-09847-z . Accessed 2021-07-17

[61] Spanakis, E.K., Golden, S.H.: Race/ethnic difference in diabetes and diabetic complications. Current diabetes reports **13**, 814–823 (2013) https://doi.org/10.1007/s11892-013-0421-9

[62] Calders, T., Žliobaitė, I.: In: Custers, B., Calders, T., Schermer, B., Zarsky, T. (eds.) Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures, pp. 43–57. Springer, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-30487-3_3

[63] Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and information systems **33**(1), 1–33 (2012) https://doi.org/10.1007/s10115-011-0463-8

[64] Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data **5**(2), 153–163 (2017) https://doi.org/10.1089/big.2016.0047

[65] Zehlike, M., Hacker, P., Wiedemann, E.: Matching code and law: achieving algorithmic fairness with optimal transport. Data Mining and Knowledge Discovery **34**(1), 163–200 (2020) https://doi.org/10.2139/ssrn.3470026

[66] Calders, T., Verwer, S.: Three naive bayes approaches for discrimination-free classification. Data mining and knowledge discovery **21**, 277–292 (2010) https://doi.org/10.1007/s10618-010-0190-x

[67] Mishler, A., Kennedy, E.H., Chouldechova, A.: Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21, pp. 386–400. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3442188.3445902

[68] Sokol, K.: Fairness, accountability and transparency in artificial intelligence: A case study of logical predictive models. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. AIES '19, pp. 541–542. Association for

Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3306618.3314316

[69] Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J.M.F., Eckersley, P.: Explainable machine learning in deployment. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20, pp. 648–657. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3351095.3375624

[70] Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20, pp. 607–617. Association for Computing Machinery, Barcelona, Spain (2020). https://doi.org/10.1145/3351095.3372850

[71] Kasirzadeh, A., Smart, A.: The use and misuse of counterfactuals in ethical machine learning. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21, pp. 228–236. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3442188.3445886

[72] Gomez, O., Holter, S., Yuan, J., Bertini, E.: Vice: Visual counterfactual explanations for machine learning models. In: Proceedings of the 25th International Conference on Intelligent User Interfaces. IUI '20, pp. 531–535. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3377325.3377536

[73] Sharma, S., Henderson, J., Ghosh, J.: Certifai: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. AIES '20, pp. 166–172. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3375627.3375812

[74] Swaminathan, A., Joachims, T.: Batch learning from logged bandit feedback through counterfactual risk minimization. J. Mach. Learn. Res. **16**(1), 1731–1755 (2015). https://api.semanticscholar.org/CorpusID:7297845

[75] Ramsahai, R.R.: Causal bounds and observable constraints for non-deterministic models. J. Mach. Learn. Res. **13**(1), 829–848 (2012). https://dl.acm.org/doi/10.5555/2503308.2188414

[76] Russell, C.: Efficient search for diverse coherent explanations. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT* '19, pp. 20–28. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3287560.3287569

[77] Li, Y., Ning, Y., Liu, R., Wu, Y., Hui Wang, W.: Fairness of classification using

users' social relationships in online peer-to-peer lending. In: Companion Proceedings of the Web Conference 2020. WWW '20, pp. 733–742. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3366424.3383557

[78] Tavakol, M.: Fair classification with counterfactual learning. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '20, pp. 2073–2076. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3397271.3401291

[79] Rosenfeld, N., Mansour, Y., Yom-Tov, E.: Predicting counterfactuals from large historical data and small randomized trials. In: Proceedings of the 26th International Conference on World Wide Web Companion. WWW '17 Companion, pp. 602–609. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2017). https://doi.org/10.1145/3041021.3054190

[80] Huan, W., Wu, Y., Zhang, L., Wu, X.: Fairness through equality of effort. In: Companion Proceedings of the Web Conference 2020. WWW '20, pp. 743–751. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3366424.3383558

[81] Coston, A., Mishler, A., Kennedy, E.H., Chouldechova, A.: Counterfactual risk assessments, evaluation, and fairness. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20, pp. 582–593. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3351095.3372851

[82] Amjad, M., Shah, D., Shen, D.: Robust synthetic control. J. Mach. Learn. Res. **19**(1), 802–852 (2018). http://jmlr.org/papers/v19/17-777.html

[83] Zeng, S., Bayir, M.A., Pfeiffer, J.J., Charles, D., Kiciman, E.: Causal transfer random forest: Combining logged data and randomized experiments for robust prediction. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining. WSDM '21, pp. 211–219. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3437963.3441722

[84] Black, E., Fredrikson, M.: Leave-one-out unfairness. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 285–295 (2021). https://doi.org/10.1145/3442188.3445894

[85] Tolan, S., Miron, M., Gómez, E., Castillo, C.: Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in catalonia. In: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law. ICAIL '19, pp. 83–92. Association for Computing Machinery,

New York, NY, USA (2019). https://doi.org/10.1145/3322640.3326705

[86] Dmitriev, P., Gupta, S., Kim, D.W., Vaz, G.: A dirty dozen: Twelve common metric interpretation pitfalls in online controlled experiments. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '17, pp. 1427–1436. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3097983.3098024

[87] De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., Kalai, A.T.: Bias in bios: A case study of semantic representation bias in a high-stakes setting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT* '19, pp. 120–128. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3287560.3287572

[88] Baeza-Yates, R.: Bias on the web. Commun. ACM **61**(6), 54–61 (2018) https://doi.org/10.1145/3209581

[89] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N.: Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '15, pp. 1721–1730. Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2783258.2788613

[90] Lipton, Z.C.: The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue **16**(3), 31–57 (2018) https://doi.org/10.1145/3236386.3241340

[91] Yang, K., Qinami, K., Fei-Fei, L., Deng, J., Russakovsky, O.: Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20, pp. 547–558. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3351095.3375709

[92] Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. AIES '18, pp. 335–340. Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3278721.3278779

[93] Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '15, pp. 259–268. Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2783258.2783311

[94] Salimi, B., Rodriguez, L., Howe, B., Suciu, D.: Interventional fairness: Causal database repair for algorithmic fairness. In: Proceedings of the 2019 International Conference on Management of Data. SIGMOD '19, pp. 793–810. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3299869.3319901

[95] Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Friedler, S.A., Wilson, C. (eds.) Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Proceedings of Machine Learning Research (PMLR), vol. 81, pp. 77–91 (2018). https://proceedings.mlr.press/v81/buolamwini18a.html

[96] Wang, H., Ustun, B., Pin Calmon, F.: Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In: International Conference on Machine Learning (2019). https://api.semanticscholar.org/CorpusID:59413891

[97] Zafar, M.B., Valera, I., Gomez-Rodriguez, M., Gummadi, K.P.: Fairness constraints: Mechanisms for fair classification. In: International Conference on Artificial Intelligence and Statistics (2015). https://api.semanticscholar.org/CorpusID:8529258

[98] Zemel, R.S., Wu, L.Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International Conference on Machine Learning (2013). https://api.semanticscholar.org/CorpusID:490669

[99] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. ITCS '12, pp. 214–226. Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2090236.2090255

[100] Madras, D., Creager, E., Pitassi, T., Zemel, R.S.: Learning adversarially fair and transferable representations, vol. abs/1802.06309 (2018). https://api.semanticscholar.org/CorpusID:3419504

[101] Gordaliza, P., Barrio, E., Gamboa, F., Loubes, J.-M.: Obtaining fairness using optimal transport theory. In: International Conference on Machine Learning (2018). https://api.semanticscholar.org/CorpusID:67780032

[102] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H.M.: A reductions approach to fair classification, vol. abs/1803.02453 (2018). https://api.semanticscholar.org/CorpusID:4725675

[103] Suresh, H., Guttag, J.V.: A framework for understanding unintended consequences of machine learning, vol. abs/1901.10002 (2019). https://api.semanticscholar.org/CorpusID:59336269

[104] Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., Sculley, D.: No classification without representation: Assessing geodiversity issues in open data sets for the developing world. arXiv (2017). https://api.semanticscholar.org/CorpusID:26262581

[105] Friedler, S.A., Scheidegger, C.E., Venkatasubramanian, S.: On the (im)possibility of fairness, vol. abs/1609.07236 (2016). https://api.semanticscholar.org/CorpusID:263792047

[106] Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv: Machine Learning (2017). https://api.semanticscholar.org/CorpusID:11319376

[107] Datta, A., Tschantz, M.C., Datta, A.: Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. arXiv preprint arXiv:1408.6491 (2014) https://doi.org/10.48550/arXiv.1408.6491

[108] Feng, R., Yang, Y., Lyu, Y., Tan, C., Sun, Y., Wang, C.: Learning fair representations via an adversarial framework. arXiv preprint arXiv:1904.13341 (2019) https://doi.org/10.48550/arXiv.1904.13341

[109] Lum, K., Johndrow, J.: A statistical framework for fair predictive algorithms. arXiv preprint arXiv:1610.08077 (2016) https://doi.org/10.48550/arXiv.1610.08077

[110] Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. In: Ethics of Data and Analytics, pp. 254–264 (2016)

[111] Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. Ethics of Data and Analytics: Concepts and Cases, 254 (2022) https://doi.org/10.1201/9781003278290

[112] Xu, D., Wu, Y., Yuan, S., Zhang, L., Wu, X.: Achieving causal fairness through generative adversarial networks. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, pp. 1452–1458 (2019). https://doi.org/10.24963/ijcai.2019/201

[113] Dwork, C., Immorlica, N., Kalai, A.T., Leiserson, M.: Decoupled classifiers for fair and efficient machine learning. arXiv preprint (2017) https://doi.org/10.48550/arXiv.1707.06613

[114] Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019) https://doi.org/10.1126/science.aax2342

[115] Dressel, J., Farid, H.: The accuracy, fairness, and limits of predicting recidivism. Science advances **4**(1), 5580 (2018) https://doi.org/10.1126/sciadv.aao5580

[116] Thomas, P.S., Silva, B., Barto, A.G., Giguere, S., Brun, Y., Brunskill, E.: Preventing undesirable behavior of intelligent machines. Science **366**(6468), 999–1004 (2019) https://doi.org/10.1126/science.aag3311

[117] Skeem, J.L., Lowenkamp, C.T.: Risk, race, and recidivism: Predictive bias and disparate impact. Criminology **54**(4), 680–712 (2016) https://doi.org/10.1111/1745-9125.12123

[118] Danks, D., London, A.J.: Algorithmic bias in autonomous systems. In: Ijcai, vol. 17, pp. 4691–4697 (2017). https://api.semanticscholar.org/CorpusID:33799296

[119] Stevenson, M.: Assessing risk assessment in action. LSN: Criminal Procedure (Topic) (2018) https://doi.org/10.2139/ssrn.3016088

[120] Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence **1**(5), 206–215 (2019) https://doi.org/10.1038/s42256-019-0048-x

[121] Cowgill, B., Tucker, C.: Algorithmic bias: A counterfactual perspective. NSF Trustworthy Algorithms (2017). https://api.semanticscholar.org/CorpusID:53961090

[122] Russell, C., Kusner, M.J., Loftus, J.R., Silva, R.: When worlds collide: Integrating different counterfactual assumptions in fairness. In: Neural Information Processing Systems, vol. 30 (2017). https://api.semanticscholar.org/CorpusID:3558923

[123] Bolukbasi, T., Chang, K.-W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems **29** (2016) https://doi.org/10.48550/arXiv.1607.06520

[124] Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., Varshney, K.R.: Optimized pre-processing for discrimination prevention. Advances in neural information processing systems **30** (2017) https://doi.org/10.48550/arXiv.1704.03354

[125] Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016) https://doi.org/10.48550/arXiv.1610.02413

[126] Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., Weinberger, K.Q.: On fairness and calibration. Advances in neural information processing systems **30** (2017) https://doi.org/10.48550/arXiv.1709.02012

[127] Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. Advances in neural information processing systems **30** (2017) https://doi.org/10.48550/

arXiv.1703.06856

[128] Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., Schölkopf, B.: Avoiding discrimination through causal reasoning. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17, pp. 656–666. Curran Associates Inc., Red Hook, NY, USA (2017). https://dl.acm.org/doi/10.5555/3294771.3294834

[129] Chiappa, S.: Path-specific counterfactual fairness. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 7801–7808 (2019). https://doi.org/10.1609/aaai.v33i01.33017801

[130] Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018). https://doi.org/10.1609/aaai.v32i1.11491

[131] Maddox, A.: Introduction to Statistical Methods (2016). https://he.kendallhunt.com/product/introduction-statistical-methods Accessed 2023-05-06

[132] Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. arXiv (2016) https://doi.org/10.48550/arXiv.1609.05807

[133] Di Stefano, P.G., Hickey, J.M., Vasileiou, V.: Counterfactual fairness: removing direct effects through regularization. arXiv (2020) https://doi.org/10.48550/arXiv.2002.10774

[134] Valera, I., Singla, A., Gomez Rodriguez, M.: Enhancing the accuracy and fairness of human decision making. Advances in Neural Information Processing Systems **31** (2018). https://dl.acm.org/doi/10.5555/3326943.3327106

[135] Nabi, R., Shpitser, I.: Fair inference on outcomes. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018). https://doi.org/10.1609/aaai.v32i1.11553

[136] Xu, D., Yuan, S., Zhang, L., Wu, X.: Fairgan: Fairness-aware generative adversarial networks. In: 2018 IEEE International Conference on Big Data (Big Data), pp. 570–575 (2018). https://doi.org/10.1109/bigdata.2018.8622525 . IEEE

[137] Devlin, S., Yliniemi, L., Kudenko, D., Tumer, K.: Potential-based difference rewards for multiagent reinforcement learning. In: Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems. AAMAS '14, pp. 165–172. International Foundation for Autonomous Agents and Multiagent Systems, Paris, France (2014). https://dl.acm.org/doi/10.5555/2615731.2615761

[138] Colby, M.K., Kharaghani, S., HolmesParker, C., Tumer, K.: Counterfactual

Exploration for Improving Multiagent Learning. In: Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems. AAMAS '15, pp. 171–179. International Foundation for Autonomous Agents and Multiagent Systems, Istanbul, Turkey (2015). https://api.semanticscholar.org/CorpusID:1379784

[139] Athey, S.: Machine Learning and Causal Inference for Policy Evaluation. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery And Data Mining. KDD '15, pp. 5–6. Association for Computing Machinery, Sydney, NSW, Australia (2015). https://doi.org/10.1145/2783258.2785466

[140] Guidotti, R., Monreale, A., Spinnato, F., Pedreschi, D., Giannotti, F.: Explaining Any Time Series Classifier. In: 2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI), pp. 167–176 (2020). https://doi.org/10.1109/CogMI50398.2020.00029

[141] Cheng, F., Ming, Y., Qu, H.: DECE: Decision Explorer with Counterfactual Explanations for Machine Learning Models. IEEE Transactions on Visualization and Computer Graphics **27**(2), 1438–1447 (2021) https://doi.org/10.1109/TVCG.2020.3030342

[142] Barocas, S., Selbst, A.D., Raghavan, M.: The hidden assumptions behind counterfactual explanations and principal reasons. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 80–89. ACM, Barcelona Spain (2020). https://doi.org/10.1145/3351095.3372830

[143] Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech. **31**, 841 (2017) https://doi.org/10.48550/arXiv.1711.00399

[144] Samadi, S., Tantipongpipat, U., Morgenstern, J., Singh, M., Vempala, S.: The price of fair pca: One extra dimension. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. NIPS'18, pp. 10999–11010. Curran Associates Inc., Red Hook, NY, USA (2018). https://dl.acm.org/doi/10.5555/3327546.3327755

[145] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '17, pp. 797–806. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3097983.3098095 . https://doi.org/10.1145/3097983.3098095

[146] Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research **50**(1), 3–44 (2021) https://doi.org/10.1177/0049124118782533

[147] Chouldechova, A., Roth, A.: The frontiers of fairness in machine learning. arXiv preprint arXiv:1810.08810 (2018) https://doi.org/10.48550/arXiv.1810.08810

[148] Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S.: Human Decisions and Machine Predictions*. The Quarterly Journal of Economics **133**(1), 237–293 (2017) https://doi.org/10.1093/qje/qjx032

[149] Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., Nagar, S., Ramamurthy, K.N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y.: Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development **63**(4/5), 4–1415 (2019) https://doi.org/10.1147/JRD.2019.2942287

[150] Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16, pp. 1135–1144. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939778

[151] Wedding, D., Professor, P.: Unit 02: HELOC (Bingo Bonus Problem). Kaggle (2015). https://kaggle.com/competitions/heloc

[152] Becker, B., Kohavi, R.: Adult. UCI Machine Learning Repository (1996). https://archive.ics.uci.edu/dataset/2/adult

[153] Hofmann, H.: Statlog (German Credit Data). UCI Machine Learning Repository (1994). https://doi.org/10.24432/C5NC77

[154] Lending Club: Institutional Investing Resources. https://www.lendingclub.com/investing/investor-education

[155] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., Johannes, R.S.: Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitu. Proceedings Symposium on Computer Applications in Medical Care, 261–265 (1988). Accessed 2021-07-10

[156] Mjkistler, Locar, R., Lempel, R., RoySassonOB, R., Cukierski, W.: Outbrain Click Prediction. Kaggle (2016). https://kaggle.com/competitions/outbrain-click-prediction

[157] Howard, A., Chiu, A., McDonald, M., Msla, Kan, W., Yianchen: WSDM - KKBox's Music Recommendation Challenge. Kaggle (2017). https://kaggle.com/competitions/kkbox-music-recommendation-challenge

[158] Yuan, B., Liu, Y., Hsia, J.-Y., Dong, Z., Lin, C.-J.: Unbiased ad click prediction for position-aware advertising systems. In: Fourteenth ACM Conference

on Recommender Systems. RecSys '20, pp. 368–377. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3383313.3412241

[159] Consumer Financial Protection Bureau: Historic HMDA Data. https://www.consumerfinance.gov/data-research/hmda/historic-data/

[160] Petersen, K., Vakkalanka, S., Kuzniarz, L.: Guidelines for conducting systematic mapping studies in software engineering: An update. Information and Software Technology **64**, 1–18 (2015) https://doi.org/10.1016/j.infsof.2015.03.007

[161] Gonçales, L., Farias, K., Silva, B., Fessler, J.: Measuring the cognitive load of software developers: A systematic mapping study. In: 2019 IEEE/ACM 27th International Conference on Program Comprehension (ICPC), pp. 42–52 (2019). https://doi.org/10.1109/ICPC.2019.00018