# HYBRID FAIRNESS FOR FAIR DECISION MAKING

**Erdem KUŞ**

**MAY 2022**

# HYBRID FAIRNESS FOR FAIR DECISION MAKING

**A THESIS SUBMITTED TO THE**

**GRADUATE SCHOOL**

**OF**

**BAHÇEŞEHİR UNIVERSITY**

**ERDEM KUŞ**

**IN PARTIAL FULFILLMENT OF THE**

**REQUIREMENTS FOR**

**THE DEGREE OF MASTER OF SCIENCE**

**IN THE DEPARTMENT OF ARTIFICIAL INTELLIGENCE**

**MAY 2022**

**T.C.**

**BAHÇEŞEHİR UNIVERSITY**

**GRADUATE SCHOOL**

...../...../.....

## MASTER THESIS APPROVAL FORM

| | |
|---|---|
| **Program Name:** | |
| **Student's Name and Surname:** | |
| **Name of The Thesis:** | |
| **Thesis Defense Date** | |

This thesis has been approved by the Graduate School which has fulfilled the necessary conditions as Master thesis.

**Prof. Dr. Ahmet ÖNCÜ**

**Institute Director**

This thesis was read by us, quality and content as a Master's thesis has been seen and accepted as sufficient.

| | **Title/Name** | **Signature** |
|---|---|---|
| **Thesis Advisor's** | | |
| **2. Member's** | | |
| **3. Member's** | | |

**I hereby declare that all information in this document has been obtained and presentedin accordance with academic rules and ethical conduct. I also declare that, as requiredby these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.**

Name, Surname :

Signature :

**ABSTRACT**

HYBRID FAIRNESS FOR FAIR DECISION MAKING

Erdem KUŞ

Artificial Intelligence Master Program

Supervisor: Asst. Prof. Dr. Görkem KAR

May 2022, 102 pages

In recent years, machine learning (ML) has achieved impressive progress and proved that it could outperform humans in many applications. As a consequence of this performance, machine learning systems are increasingly being used in real-world problems, and this led society to view artificial intelligence more critically. One of the biggest factors in society's skepticism towards machine learning-supported systems is that the models have biased outputs that may offend a certain group. With the increase in such biased decisions, fairness has taken its place in the world of artificial intelligence, as a metric that is at least as important as accuracy and should be taken into account.

With the introduction of fairness into artificial intelligence systems, the concept of fair artificial intelligence, which covers the whole of the studies aimed at preventing biased decisions, has emerged. Fair AI includes studies that aim to prevent biased decisions of the model, mostly by using group fairness metrics. However, since group fairness is inversely proportional to accuracy and causes unjust results to successful individuals in priority groups, is insufficient in real-world systems. In addition, another problem related to fairness is that fairness metrics are tried to be proven only with an inter-group statistical metric, and their explainability is ignored. In this study, we introduce the concept of hybrid fairness, which is the combination of group fairness and, individual fairness, as a solution to the problems mentioned, and we present the outputs of this concept with both statistical and explainability in this work.

**Keywords**: Fair Artificial Intelligence, Group Fairness, Individual Fairness, Explainable Artificial Intelligence,

# ÖZ

ADİL KARAR VERME SİSTEMLERİ İÇİN HİBRİT ADALET

Erdem KUŞ

Yapay Zeka Yüksek Lisans Programı

Tez Danışmanı: Dr. Öğr. Üyesi Görkem KAR

Mayıs 2022, 102 sayfa

Son yıllarda makine öğrenimi (ML) etkileyici bir ilerleme kaydetti ve birçok uygulamada insanlardan daha iyi performans gösterebileceğini kanıtladı. Bu performansın bir sonucu olarak, makine öğrenme sistemleri gerçek dünya problemlerinde giderek daha fazla kullanılmaktadır ve bu da toplumu yapay zekaya daha eleştirel bakmaya yöneltmiştir. Toplumun makine öğrenimi destekli sistemlere yönelik şüpheciliğindeki en büyük etkenlerden biri, modellerin belirli bir grubu rahatsız edebilecek önyargılı çıktılara sahip olmasıdır. Bu tür önyargılı kararların artmasıyla birlikte adalet, en az doğruluk kadar önemli ve dikkate alınması gereken bir metrik olarak yapay zeka dünyasında yerini almıştır.

Adaletin yapay zeka sistemlerine girmesiyle birlikte önyargılı kararların önlenmesine yönelik çalışmaların bütününü kapsayan adil yapay zeka kavramı ortaya çıkmıştır. Fair AI, çoğunlukla grup adaleti metriklerini kullanarak modelin önyargılı kararlarını önlemeyi amaçlayan çalışmaları içerir. Ancak, grup adaleti modelin başarısı ters orantılı olduğundan ve öncelikli gruplardaki başarılı bireylere adil olmayan sonuçlara neden olabileceğinden, gerçek dünya sistemlerinde yetersizdir. Ayrıca adaletle ilgili bir diğer sorun da adalet metriklerinin sadece gruplar arası istatistiksel bir metrik ile kanıtlanmaya çalışılması ve açıklanabilirliğinin göz ardı edilmesidir. Bu çalışmada, bahsedilen sorunlara çözüm olarak grup adaleti ile bireysel adaletin birleşimi olan hibrit adalet kavramını tanıtıyoruz ve bu kavramın çıktılarını bu çalışmada hem istatistiksel hem de açıklanabilir olarak sunuyoruz.

**Anahtar Kelimeler**: Adil Yapay Zeka, Grup Adaleti, Bireysel Adalet, Açıklanabilir Yapay Zeka

***To my mother***

## ACKNOWLEDGMENTS

First of all, I would like to thank my thesis advisor, Asst. Prof. Dr. Görkem KAR for his support and trust in me during the thesis preparation process. I'm very grateful for his support, insight, and invaluable help during the preparation of this thesis.

I would also like to thank my mother, Gülgün KUŞ, for their great support throughout her life. I am grateful for the vision and moral values she has given me and for allowing me to be me.

I would like to thank Kenan KIRATLI, my manager at Nokia, for his understanding and support throughout this process. I would also like to thank my company and my teammates.

Finally, I would like to thank William Sean Kennedy, the leader of the Nokia Bell Labs Artificial Intelligence Lab, and Benoit Drooghaag, an artificial intelligence engineer at Nokia, for their guidance and unconditional support.

Istanbul, 2022 Erdem KUŞ

## TABLE OF CONTENTS

## LIST OF TABLES

TABLES

## TABLE OF FIGURES

FIGURES

## LIST OF ABBREVIATIONS

| | |
|---|---|
| AI | Artificial Intelligence |
| AAAI | Association for the Advancement of Artificial Intelligence |
| ACCV | Asian Conference on Computer Vision |
| ACM | Association for Computing Machinery |
| AMD | Age Related Macular Degeneration |
| ANN | Artificial Neural Networks |
| BERT | Bidirectional Encoder Representations from Transformers |
| DARPA | Defense Advanced Research Projects Agency |
| DBSCAN | Density-Based Spatial Clustering of Applications with Noise |
| CASP | Critical Assessment of protein Structure Prediction |
| CNN | Convolutional Neural Network |
| DEX | Deep EXpectation Network |
| COMPAS | Correctional Offender Management Profiling for Alternative Sanctions |
| ECLAT | Equivalence Class Transformation |
| GAN | Generative Adversarial Networks |
| GDT | Global Distance Test |
| HGR | Hirschfeld-Gebelein-Renyl |
| IEEE | Institute of Electrical and Electronics Engineers |
| INLP | Iterative Null-space Projection |
| IoT | Internet of Things |
| ISO | International Organization for Standardization |
| IW | Instance Re-weighting |
| JAMA | Journal of the American Medical Association |
| LDAM | Label-Distribution-Aware Margin |
| LSTM | Long Short-Term Memory |
| ML | Machine Learning |
| MLP | Multi-Layer Perception |
| NIPS | Conference and Workshop on Neural Information Processing Systems |
| NLP | Natural Language Processing |

| | |
|---|---|
| OPTICS | Ordering Points to Identify the Clustering Structure |
| PAC | Probably Approximately Correct |
| ResNet | Residual Neural Network |
| RTX | Ray Tracing Texel eXtreme |
| SLR | Systematic Literature Review |
| SVM | Support Vector Machine |
| VAE | Variational Autoencoder |
| VGG | Visual Geometry Group |
| XAI | Explainable Artificial Intelligence |

# Chapter 1: Introduction

Machine learning (ML) achieved impressive progress in several areas including games (Vinyals et al., 2019), computer vision (Xiao et al., 2020), neuroinformatics (Spape et al., 2021), and so on. Developments in machine learning have allowed the creation of scalable, fast, and powerful systems and the use of these systems in real life. With machine learning starting to make decisions for people in daily life such as credit lending (Dwork, Hardt, Pitassi, Reingold, & Zemel, 2012), criminal justice (Barry-Jester, Casselman, & Goldstein, 2015), and policing (Rudin, 2013), society started to look more critically at the decisions made with machine learning. Decisions made by intelligent algorithms can have long-term effects on the human's daily life and notably may affect particular individuals or social groups in a negative way (“Machine Bias — ProPublica,” n.d.).

## 1.1 Theorical Framework

Different studies (Caliskan, Bryson, & Narayanan, 2017; Chouldechova, 2017; Obermeyer, Powers, Vogeli, & Mullainathan, 2019) also have shown us that current machine learning models might reflect prejudices from protected feature bias which can be categorized as such as gender, sexual preference, age, race, and so on. We will present the problems caused by some biased machine learning systems from different fields in the next sections.

To address this problem, fairness has become a metric for artificial intelligence models and different studies have been developed in the light of different fairness metrics. Fair AI includes three methodologies to reduce bias and improve fairness in AI models: 1) pre-processing, 2) in-process, and 3) post-processing.

In the preprocessing approach (Celis, Keswani, & Vishnoi, 2020; del Barrio, Gamboa, Gordaliza, & Loubes, 2019; Louizos, Swersky, Li, Welling, & Zemel, 2016; Zemel, Wu, Swersky, Pitassi, & Dwork, 2013), the aim can be defined as training artificial intelligence models with less bias by reducing the bias in the data set. This can be done by learning a new representation or adjusting the representation rates of

protected groups to achieve the selected metric (group or individual). This adjusted dataset is convenient to be used for any machine learning or deep learning technique for any task such as classification, and decision making. The main advantage of the approach is, that there is no need for fairness modification on the machine learning algorithm. On the other hand, performance on accuracy and fairness metrics is not as good as the other approaches.

The in-processing approach consists of the techniques to mitigate bias in the training process. Most of the methods in fair artificial intelligence belong to this approach (Madras, Creager, Pitassi, & Zemel, 2019; Noroozi, Bahaadini, Sheikhi, Mojab, & Yu, 2019; Ravichandran et al., 2020; Zafar, Valera, Gomez Rodriguez, & Gummadi, 2017). Algorithms in this category mostly use regularization which is derived from existing or proposed fairness metrics. The main advantage of these approaches is the robust and flexible handling of the trade-off between fairness and accuracy. However, most of the approaches are task-specific therefore generalization of the models is suspicious for this approach.

Lastly, the post-processing approach includes the proposed techniques in order to mitigate the bias of the output instead of training data or training phase (Hardt, Price, & Srebro, 2016; Lohia et al., 2019; Petersen, Mukherjee, Sun, & Yurochkin, n.d.). Algorithms in this approach control the output for protected groups by using fairness metrics. The main advantage of the approach is there is no need to change the machine learning model nevertheless it is not robust and flexible in accuracy fairness trade-off.

It is quite difficult to cover the whole concept of fair artificial intelligence, as different approaches can cause different problems, as outlined in Table 1.

Table 1

*Potential Problems in Fair Machine Learning Approaches*

| Fairness Approach | Potential Problem |
|---|---|
| Pre-processing | Ensuring fairness in data without protected features can be a problem.<br><br>Performance on fairness and accuracy is quite difficult. |

Table 1 (continued)

| Fairness Approach | Potential Problem |
|---|---|
| | Only suitable for specific fairness metrics such as demographic parity and equalized odds due to missing label information. |
| | Only suitable for task-specific examples, therefore generalization is a problem. |
| In-processing | Modifying the machine learning model might not be possible. |
| | Adversarial attack might be a problem for this approach. |
| Post-processing | Machine learning output is required; therefore, it can take time. |
| | Performance on fairness and accuracy is quite difficult |

The fairness metrics used in these approaches are divided into two groups as individual fairness and group fairness. In group fairness, the main idea is individuals should be treated the same at the group level. In other words, groups should have the same statistics, such as TPR (True Positive Rate) and/or False Positive Rate (FPR). It is the most used fairness type in the fair machine learning approach because most of the mistreatment is at group-level. In individual fairness, similar individuals should get similar rewards. Both metrics will be investigated deeply in further sections.

### 1.2 Statement of the Problem

In the systematic literature review, we have done, we have presented that group fairness is used to a great extent in the studies that we have drawn attention and examined in depth. One of the reasons for this is that prejudice is based on protected attributes and unfair results are obtained for the group of individuals who inductively have these attributes. Group-based fairness is based on the equality of group-based ratios or combinations of metrics such as TPR, FPR, TNR, FNR. However, since it is impossible to provide all of these ratios at the same time due to unbalanced groups, noise or mislabeling, different group fairness metrics have used these ratios in different ways. The advantages and disadvantages of these ratios are shown in Figure 1. As can

be seen in the table, the biggest disadvantage of group fairness metrics is that it creates unfair advantage or disadvantage on individuals. Our main motivation in this research, is that the work done to eliminate the injustice of underprivileged groups leads to individual injustices. To solve this problem, we define hybrid fairness, which can be defined as a combination of both individual and group fairness.

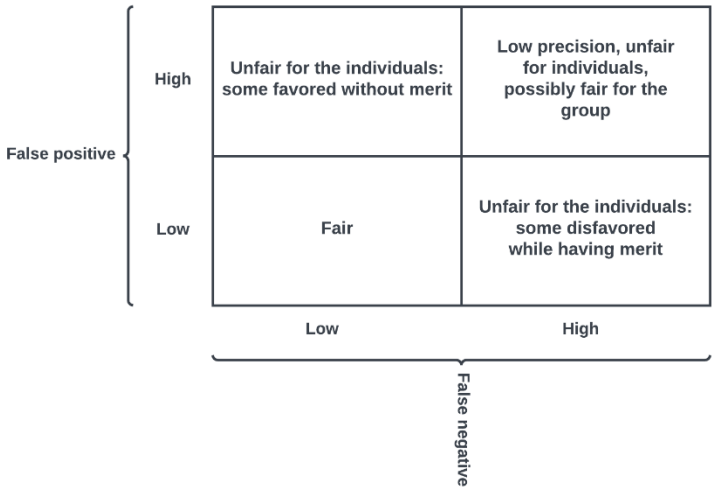


*Figure 1.* Advantages and disadvantages of ratios

### 1.3 Purpose of the Study

In this work, we aim to define a new metric to mitigate the disadvantages of the group and individual fairness notions. For this reason, we define fairness as a constraint satisfaction problem as it can be defined as minimizing group and individual based disadvantages. Since the use of group-based fairness in real-time systems may cause individuals in groups with limited quotas, which are seen as advantageous in cases such as scholarship application, school admission and loan application, to a disadvantageous situation, fairness between the group and the individual needs to be balanced. In addition, developing machine learning models by basing fairness only on statistical aspect such as being elected and unelected may lead to statistical uncertainty due to the different behavior of different fairness metrics and an ontological uncertainty of fairness for the end user. For this reason, we present the hybrid fairness that we have stated within the scope of explainability. In this way, we hope that, in addition to our main objectives, the evaluation of group and individual fairness

together will encourage studies on which group and individual fairness metrics give more positive or more negative results with each other.

### 1.4 Hypotheses/Research Questions

We presented this study by seeking answers to the questions stated in this study.

- What is fair artificial intelligence?
- What is the importance of fair AI?
- What are the metrics used in fair AI and what are the disadvantages of these metrics?
- What are the concepts that should be evaluated together with fairness?

Our first question shows the motivation of our literature review, our second question shows the possible impact of the work we will do, our third question shows the main objective of our study and our last question shows the scope of explainability of our work.

### 1.5 Significance of the Study

The fact that group fairness causes successful individuals in groups to get negative results has been our starting point in this research. As important as group fairness is, it is equally important for an individual to get a fair result. Therefore, we present the hybrid metric defined as the combination of group and individual fairness. With this metric, the balance between group fairness and individual fairness will be achieved more precisely.

Our second contribution to this work is that these fairness metrics are only based on a group or individual statistical calculation, which limits their explainability. Although tree-based fair algorithms (Ranzato, Urban, & Zanella, 2021; W. Zhang & Bifet, 2020; W. Zhang & Ntoutsi, 2019) are high explainable in nature, it is necessary to generalize the explainability for black-box models with low explainability.

## 1.6 Definitions

Artificial intelligence models can be defined as a producer of the function which calculates desirable output by using different learning methodologies for different problem types. Performance of the models is measured by using statistical metrics such as accuracy, binary cross-entropy, and so on, between the model's output and desirable output. However, the increased number of using artificial intelligence models shows us that output expectation cannot be a sole metric for artificial intelligence models therefore fairness should be considered for real-time decision-risk applications. In this section, we will provide key terms with their definitions and criticism with the help of related studies regarding these key terms.

**1.6.1 Artificial intelligence.** Artificial intelligence term has been proposed by John McCarthy (Mccarthy, 2007) as "It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods biologically observable.". Over the decades, several algorithms and learning strategies have been developed in light of this definition. Although there are learning models derived from these learning methods or produced specific to certain conditions, artificial intelligence is mainly based on 3 learning strategies.

***1.6.1.1 Supervised learning***. The supervised learning comprises classification and regression algorithms that take a feature vector and the label/value as an input for each data segment to produce a function so that model can learn how to classify or predict when a new data is given.

In the classification task, machine learning models are used to assign test data into specific categories accurately. It aims to create a pattern that determines which class a new data belongs to by using the test data and the classes it belongs to. Classification trees, support vector machines (SVMs), random forests, artificial neural networks (ANNs), or other algorithms are convenient for this task.

In the regression, the main idea is to find the correct correlation between input features and outputs. It is mostly used for forecasting tasks. Regression algorithms include linear regression, logistic regression, decision trees, Bayesian networks, and ANNs.

There are some difficulties/challenges in supervised learning:

- Redundant feature can decrease the performance of the model therefore feature engineering is important.
- Adversarial attacks can cause serious problems therefore security and privacy of the model should be deeply inspected.
- Noise or anomalous data can cause unreliable outputs therefore good data processing is required.

***1.6.1.2 Unsupervised learning.*** Unlike supervised learning, unsupervised learning doesn't need labels to learn a function. Instead of labels, unsupervised learning methodologies use a statistical or probabilistic measurement of the data points such as density, distance, and similarity metrics. Clustering and assignment problems are the main problems of this learning methodology.

In clustering, unsupervised learning is used for finding a structure or pattern to agglomerate data points in a collection of uncategorized data. K-means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Ordering Points To Identify the Clustering Structure (OPTICS), ANN, and so on.

In the association, the aim is to discover relations between variables in large databases. APriori, Equivalence Class Transformation (ECLAT), and FP-growth algorithms are used to perform this task.

***1.6.1.3 Reinforcement learning.*** Reinforcement learning is different from that of both supervised and unsupervised learning. Reinforcement learning consists of agents, environments, states, actions, and rewards instead of feature vectors and labels. The reinforcement learning agent chooses what to do and reinforcement learning algorithms aim to maximize the reward of an agent's action in a state. Reinforcement

learning is mostly applied in games such as chess, go, or computer games but it has started to find wider implementation areas.

Besides these three learning methodologies, hybrid or derived learning methodologies such as semi-supervised learning, self-supervised learning, transductive learning, and so on have been proposed by researchers to mitigate problems in three main learning methodologies, improve learning performance, and solve specific tasks. The diversity in learning strategies has created an opportunity to embed AI models into real-world applications, and these models far outperform humans in complex tasks in many areas can be seen in Table 2.

Table 2

*Some of the Machine Learning Use Cases*

| Use Case | Approach | Reference |
|---|---|---|
| Games | Deep Mind developed an agent named Alpha star using reinforcement learning was rated at Grandmaster level (highest rank in the game) for all three StarCraft races (Zerg, Protoss, Human) and placed higher than 99.8% of officially ranked human players | (Vinyals et al., 2019) |
| Healthcare | Artificial intelligence improves illness detection notably in cancer by using classification and computer vision methodologies. In most cases (around 70%) of lung cancer, the disease is detected in later stages with human expertise however survival rate is too low for later stages. AI methodologies detect the disease in earlier stages therefore mortality is reduced by 20–30% | (Svoboda, 2020) |

Table 2 (continued)

| Use Case | Approach | Reference |
|---|---|---|
| Biology | AI models are widely used in biological classification problems however the most striking example in the biology field is an AI solution to a 50-year-old grand challenge in biology named as AlphaFold from DeepMind. This problem lies in protein structure prediction and their latest AlphaFold system achieves a median score of 92.4 GDT overall across all targets in the 14th CASP assessment. | (Tunyasuvunakool et al., 2021) |

**1.6.2 Fairness.** With the increased usage of machine learning models in real-time systems, there is a growing concern in societies about biased decisions based on protective attributes such as age, gender, sex, race, and so on. We will present some examples of biased decisions in different areas. Table 3 illustrates different examples from biased machine learning models.

Due to these problems, fairness has found its place as a criterion in the world of artificial intelligence and different fairness metrics have been proposed reflecting different perspectives. Before the concept of fairness in artificial intelligence, it is necessary to examine the term fairness in a legal and sociological context.

With the increased usage of machine learning models in real-time systems, there is a growing concern in societies about biased decisions based on protective attributes such as age, gender, sex, race, and so on. We will present some examples of biased decisions in different areas. Table 3 illustrates different examples from biased machine learning models.

Table 3

*Examples of Bias in Different Areas*

| Biased Feature | Area | Problem | Reference |
|---|---|---|---|
| Race | Face recognition | Google Photos has labeled a black couple as gorilla in 2015 | ("Google Apologises for Photos App's Racist Blunder - BBC News," n.d.) |
| Race | Recidivism | COMPAS algorithm has predicted the African-American people's re-cidivism risk than Cau-casians | ("Machine Bias — ProPublica," n.d.) |
| Age | Medicine | Medical issues such as retinal disease have bias potential for over 80 years of due to imbalanced representation in the dataset. | (Burlina et al., 2020) |
| Race & age | Recommendation systems | Minority tend to receive unfair recommendations from online services. | (Bobadilla, Lara-Cabrera, González-Prieto, & Ortega, 2021a) |
| Gender | Translation | Google Translate converts these Turkish sentences with genderless pronouns: "O bir doktor. O bir hemsire." to these English sentences: "He is a doctor. She is a nurse." | (Caliskan et al., 2017) |

Table 3 (continued)

| Biased Feature | Area | Problem | Reference |
|---|---|---|---|
| Gender | Zoology | Female koalas likely to be under-represented, and koalas higher in taller trees detected less frequently from drones | (Corcoran, Denman, & Hamilton, 2021) |

***1.6.2.1 Fairness in law and sociology***. Although fairness began to be used as a statistical measurement value only in the field of artificial intelligence, this statistical measurement continued to exist as a political, legal, and sociological phenomenon. Fairness emerges as a concept that is evaluated with the concepts of egalitarianism, consistency, discrimination, and justice. Although these concepts cause the concept of fairness to be divided into two sharp sub-concepts in artificial intelligence in the future, the main idea is based on the idea of treating people equally. Although discrimination has been defined for a long time as the practice of behavior that contradicts the concept of egalitarianism, the European Court of Human Rights states that 'following the principles which may be extracted from the legal practice of a large number of democratic States,' has held that a difference in treatment is only discriminatory when it 'has no objective and reasonable justification.' (Holmes, 2005). Since the protected features of the person are counted as an objective and reasonable justification in artificial intelligence, the expectation of fairness for groups with the same protected characteristics forms the basis of group fairness. In other words, different treatment of the unprivileged group based on protected features does not violate the concept of egalitarianism. However, since this treatment may cause unfair results to individuals who are successful in the privileged group, individual fairness, which is the basis of individual fairness, cannot be provided here. This causes the concept of fairness in law to be divided into two sub-categories as individual and group fairness in artificial intelligence. Table 4 presents the philosophical equivalents of these metrics.

Table 4

*Term Table for each Fairness Type*

| Fairness Metric | Philosophical Equivalent |
|---|---|
| Group Fairness | Egalitarianism, Anti-Discrimination |
| Individual Fairness | Aristotelian Consistency, Individual Justice |

***1.6.2.2 Fairness research in European Commission Joint Research Centre (ECJRC)***. In the report (Marion et al., 2017), the authors state one of the findings of the inequality as "people think that inequality in outcomes is more acceptable if they believe there is a high degree of equality of opportunity, that is if an individual's success is determined by hard work rather than factors beyond their control (such as race, gender, being from a wealthy family)". Another interesting finding in the report is the higher the percentage of people in a country who believe that hard work is important to progress, the lower the percentage of those who want government intervention to reduce inequality. These outputs show us that people think that the inequalities of opportunity caused by the protected features are unacceptable and contrary to fairness.

***1.6.2.3 Fairness standard in International Organization for Standardization (ISO)***. ISO 26000 (“ISO - ISO 26000 — Social Responsibility,” n.d.) is defined as the international standard to help and guide organizations effectively evaluate and address social responsibilities that are relevant and important to their mission and vision. Seven topics pertaining to social responsibility:

- Organizational governance
- Human rights
- Labor practices
- The environment
- Fair operating practices
- Consumer issues
- Community involvement and development

One of the aims of this standard is stated "assist organizations in addressing their social responsibilities while respecting cultural, societal, environmental, and

legal differences and economic development conditions". In the light of this statement, it has become a responsibility to ensure development in all segments of society by respecting people of different cultures, races, or conditions.

Fairness definitions can be categorized into three main groups: i) group fairness, ii) individual fairness, and iii) subgroup fairness.

***1.6.2.4 Group fairness***. Group fairness has been proposed first as a measurement based on statistical parity between protected groups (e.g., gender, race) in each outcome (Calders, Kamiran, & Pechenizkiy, 2009; Dwork et al., 2012; Pedreshi, Ruggieri, & Turini, 2008). According to statistical parity, groups should have an equal proportion of the positive class. Since statistical parity has been proposed, a wider spectrum of different group fairness measurements has been proposed. We will present some of the fairness measurements in group fairness.

***1.6.2.5 Group fairness notions***. In this section, we will provide the definitions and their equations of the three-most used fairness notions in group fairness.

*1.6.2.5.1 Demographic parity.* This fairness notion is also known as statistical parity in the literature and can be described as "a condition when a classifier produces outputs with equal probability for both protected and unprotected groups" (Dwork et al., 2012; Kusner, Loftus, Russell, & Silva, 2017).

$$\boldsymbol{P(P|A=0) = P(P|A=1)} \qquad 1$$

Equation 1 states that the likelihood of an outcome should be equal regardless of group. The main disadvantage of this metric is that it is sufficient to positively label random individuals to provide the metric. Therefore, the success of the individual is not considered in this metric.

*1.6.2.5.2 Equal opportunity.* Equal opportunity states that each group should get the positive outcome at equal rates (Hardt et al., 2016).

$$\boldsymbol{P}\left(\widehat{\boldsymbol{Y}} = \boldsymbol{1} \middle| \boldsymbol{A} = \boldsymbol{0}, \boldsymbol{Y} = \boldsymbol{1}\right) = \boldsymbol{P}\left(\widehat{\boldsymbol{Y}} = \boldsymbol{1}, \boldsymbol{A} = \boldsymbol{1}, \boldsymbol{Y} = \boldsymbol{1}\right) \quad 2$$

In Equation 2, the protected and unprotected groups should have equal probabilities for true positive rates (TP). However, groups with different positive ratios can cause problems in this approach.

*1.6.2.5.3 Equalized odds.* Equalized odds have been proposed by (Hardt et al., 2016) with the definition "a predictor $\hat{Y}$ satisfies equalized odds with respect to protected attribute $A$ and outcome $Y$, if $\hat{Y}$ and $A$ are independent conditional on $Y$". Equation 3 presents the formula of the definition as follows.

$$\boldsymbol{P}\{\widehat{\boldsymbol{Y}} = \boldsymbol{1} | \boldsymbol{A} = \boldsymbol{0}, \boldsymbol{Y} = \boldsymbol{y}\} = \boldsymbol{P}\{\widehat{\boldsymbol{Y}} = \boldsymbol{1} | \boldsymbol{A} = \boldsymbol{1}, \boldsymbol{Y} = \boldsymbol{y}\} \boldsymbol{y} \in \{\boldsymbol{0}, \boldsymbol{1}\} \quad 3$$

***1.6.2.6 Individual fairness.*** To the best of the author's knowledge, individual fairness has been introduced first in (Dwork et al., 2012) to mitigate the disadvantages of group (statistical) fairness on individuals. Individual fairness measurement can be defined as "similar individuals should get similar outcomes or should get treated similarly". Any measurement metric with this idea can be categorized in the individual fairness term. The main problem in this definition is the difficulty to determine which features should be used or define similar individuals. The second problem with individual fairness is similarity may reveal sensitive information about a group. Due to these reasons, individual fairness is mostly not convenient for real-world cases. Fairness through unawareness, fairness through awareness, and counterfactual fairness are some of the notions proposed in individual fairness.

***1.6.2.7 Individual fairness notions.*** In this section, we will provide the definitions and their equations of the three-most used fairness notions in individual fairness.

*1.6.2.7.1 Fairness through awareness.* (Dwork et al., 2012) has stated the definition of this notation as "any two individuals who are similar with respect to a particular task should be classified similarly". In other words, any two individuals who

are similar concerning the inverse distance for a specific task should receive similar outputs.

*1.6.2.7.2 Fairness through unawareness.* "An algorithm is fair as long as any protected attributes A are not explicitly used in the decision-making process" (Grgic-Hlača, Zafar, Gummadi, & Weller, 2016; Kusner et al., 2017). The main idea behind this notion is to prevent the effect of the protected attribute on the predictions of any individual.

*1.6.2.7.3 Counterfactual fairness*. (Kusner et al., 2017) has proposed this term that consists of two worlds namely the actual world and a counterfactual world where individuals belong to a different demographic group than the actual world, and if an individual has the same predictions in two worlds, this fairness notion is satisfied. In equation form, "Predictor $\hat{Y}$ is counterfactually fair if under any context $X = x$ and $A = a$, $P\left(\hat{Y}_{A\leftarrow a}(U) = y \middle| X = x, A = a\right) = P\left(\hat{Y}_{A\leftarrow \acute{a}}(U) = y \middle| X = x, A = a\right)$, (for all $y$ and for any value $\acute{a}$ attainable by $A$.

*1.6.2.7.4 Theil index.* The Theil index is a statistical measurement used to measure economic inequality over a population ("Theil Index," n.d.). Regional disparities are measured by a Theil entropy index, which is defined as in Equation 4.

$$\boldsymbol{T_t = T_{\alpha=1} = \frac{1}{N}\sum_{i=1}^{N}\frac{x_i}{\mu}\ln\left(\frac{x_i}{\mu}\right)} \quad 4$$

where $\mu$ is the mean income in Equation 5.

$$\boldsymbol{\mu = \frac{1}{N}\sum_{i=1}^{N} x_i} \quad 5$$

***1.6.2.8 Subgroup fairness.*** Subgroup fairness aims to combine best properties and mitigate the disadvantages of the group and individual notions of fairness (Kearns, Neel, Roth, & Wu, 2018; Kearns, Roth, Neel, & Wu, 2019). The main idea behind the subgroup fairness is, picks a statistical fairness constraint (equalizing odds, demographic parity, and so on), but then asks that this constraint holds large collection of subgroups.

**1.6.3 Explainable artificial intelligence**. Explainable artificial intelligence (XAI) is centered on the challenge of demystifying the black boxes, it also implies Responsible AI as it can help to produce transparent models. This should happen without affecting the AI model's accuracy, thus in AI in general and in ML specifically, often a trade-off must be made between accuracy and interpretability. An obvious link with the data science field arises as accuracy is closely tied to the quality and the quantity of the training data.

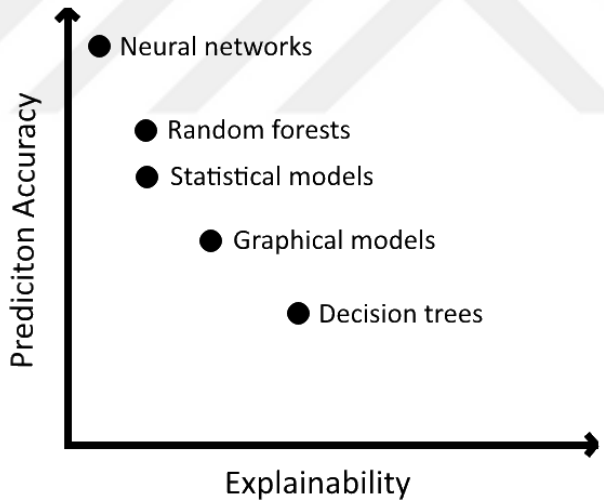


*Figure 2.* Explainability and accuracy of the different model (XAI DARPA,2016)

According to Figure 2, neural networks have the highest accuracy while having the lowest explainability. Therefore, explainability will be the key term for creating fair and accurate models. There are some explainability models which has been proposed that can be categorized into three sections 1) intrinsically interpretable methods, 2) model agnostic methods, and 3) example-based explanation methods.

***1.6.3.1 Intrinsically interpretable methods.*** Intrinsically interpretable methods are mostly focused on simple-structured models that are convenient for easy interpretability such as short decision trees and short SVMs. In other words, unlike post-hoc interpretability, intrinsically interpretable methods aim to construct self-explanatory models to satisfy interpretability directly to their structures instead of external relations.

***1.6.3.2 Model agnostic methods.*** This technique uses the extract explanations by using the black-box nature of systems (Ribeiro, Singh, & Guestrin, 2016a) namely learning an interpretable model from the predictions of the black box model and perturbing or modifying inputs and observing how the black box model (Baehrens et al., 2010; Craven & Shavlik, 1996) reacts on these inputs (Krause, Perer, & Ng, 2016; Štrumbelj & Kononenko, 2010) or a combination of two techniques (Ribeiro, Singh, & Guestrin, 2016b).

***1.6.3.3 Example-based explanation methods.*** Example-based explanation methods aim to select appropriate instances of the dataset to extract the explanation of the behavior of machine learning models. These explanation methods work only if instances can be understood well by humans therefore these methods are most suitable for computer vision cases.

The following sections are organized as follows: Section 2 provides our systematic literature review. Section 3 describes the selected datasets, our fair artificial intelligence methodology, and limitations of our methodology. Section 4 presents the results of the experiments which is performed with selected datasets and Section 5 presents the discussion, conclusion and future work.

# Chapter 2: Literature Review

Recently, there has been increasing attention on implementing fair machine learning models to mitigate bias in the systems namely decision-making, recommendation, computer vision, and natural language processing systems. In this section, we give an overview of these existing works. And also, we will share the findings of our systematic literature review on fair artificial intelligence systems.

## 2.1 Related Works

One approach to fair AI is semi-supervised approaches that take advantage of the unbiasedness of unlabeled data. One of the pioneering studies on this subject has been proposed by (Noroozi et al., 2019), which uses a semi-supervised algorithm using neural networks to improve the fairness of the decision-making process. They mostly focused on only fairness constraints in the training process which means optimizing the fairness constraint during training the classifier. According to their knowledge, the proposed methodology as SSFair is the first semi-supervised algorithm based on neural networks to make fair decisions.

In the proposed model, the aim is to optimize two main objectives, the classification accuracy, and fairness which are formulated in Equation 6.

$$\boldsymbol{J(D;\theta) = \alpha J_C(D;\theta) + (1-\alpha) J_F(D;\theta) + \beta \|\theta\|_2} \qquad 6$$

where $J_C(D;\theta)$ denotes the classification loss, and $J_F(D;\theta)$ is the fairness loss. For classification loss, they applied cross-entropy, and fairness loss is a differentiable function of group loss definitions. For model training, they have used Multi-Layer Perception (MLP) neural network with Adam optimizer and backpropagation.

One of the studies presented for fair artificial intelligence has been proposed by (H. Hu, Liu, Wang, & Lan, 2019) distributed fair learning framework for protecting the privacy of demographic data. They assume this data is privately held by a third party, which can communicate with the data center (responsible for model development) without revealing the demographic information. They propose a

principled approach to design fair learning methods under this framework, exemplify four methods, and show they consistently outperform their existing counterparts in both fairness and accuracy across three real-world data sets.

They proposed Distributed Fair Ridge Regression from Fair Ridge Regression (Calders, Karim, Kamiran, Ali, & Zhang, 2013) which minimizes squared loss on training sample, while additionally penalizing prediction disparity across demographic groups, as a regressor. Distributed Fair Ridge Regression can be formulated as in Equation 7.

$$J_{FRR}(f) = \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda . MD(f) \quad (7)$$

where $MD(f) = \frac{1}{|I_1|}\sum_{i \in I_1} f(x_i) - \frac{1}{|I_2|}\sum_{i \in I_2} f(x_i)$ is the prediction disparity.

A reductions approach to fair classification has been proposed by (Agarwal, Beygelzimer, Dudfk, Langford, & Hanna, 2018) systematic approach for achieving fairness in a binary classification setting. The key idea is to reduce fair classification to a sequence of cost-sensitive classification problems. They introduced two reductions that work for any representation of the cost-sensitive classifier and compare favorably to prior baselines on a variety of datasets. For problem reduction, the problem converts into the sequence of cost-sensitive classification problems which can be formulated as in Equation 8:

$$arg \min_{h \in H} \sum_{i=1}^{n} h(X_i)\, C_i^1 + \left(1 - h(X_i) C_i^0\right) \quad (8)$$

For fair classification algorithm, they converted problem as a saddle point problem as in Equation 9.

$$L(Q, \lambda) = \widehat{err}(Q) + \lambda^T (M\hat{\mu}Q - \hat{c}) \quad (9)$$

They have used grid search for saddle point minimization.

Some studies have used different metrics other than the one presented as the fairness metric. (Grari, Hajouji, Lamprier, & Detyniecki, 2021) point that many state-of-the-art fair machine learning models aim to learn a fair representation by capturing all relevant information except sensitive features. They proposed a new approach to mitigate bias in the representations with the Hirschfeld-Gebelein-Renyi (HGR) maximal correlation coefficient. According to them, the HGR coefficient captures more information about the non-linear dependencies with the sensitive variable compared to other metrics.

They state that fair machine learning approaches that mitigate bias by removing sensitive attributes from the training data set can cause "fairness through unawareness" (Jiahao Chen, Kallus, Mao, Svacha, & Udell, 2019) is a term that denotes the lack of required information in deep learning systems due to non-sensitive attributes might indirectly contain significant sensitive information. In their approach, they propose a neural network architecture to achieve fair representation by minimizing the HGR coefficient. For two jointly distributed random variables $U \in \mathrm{U}$ and $V \in \mathrm{V}$, the Hirschfeld-Gebelein-Rényi maximal correlation is defined as in Equation 10:

$$\boldsymbol{HGR}(\boldsymbol{U},\boldsymbol{V}) = \sup_{\boldsymbol{f}:\boldsymbol{U}\to\mathbb{R},\boldsymbol{g}\to\mathbb{R}} \boldsymbol{p}\big(\boldsymbol{f}(\boldsymbol{U}),\boldsymbol{g}(\boldsymbol{V})\big) = \sup_{\boldsymbol{f}:\boldsymbol{U}\to\mathbb{R},\boldsymbol{g}\to\mathbb{R}} \boldsymbol{E}\big(\boldsymbol{f}(\boldsymbol{U}),\boldsymbol{g}(\boldsymbol{V})\big) \qquad 10$$

The objective to achieve fair representation, find a latent representation $Z$ which both minimizes the deviation between the target $Y$ and the output prediction $\hat{Y}$, provided by a function $\phi(Z)$ and weak dependence with the sensitive attributes $S$.

### 2.2 Systematic Literature Review

While several methodologies have been proposed based on these approaches, a systematic overview of the current state-of-the-art use of fair AI is lacking. As such, we performed a Systematic Literature Review study to collect and synthesize the required data on the state-of-the-art in this field. The following research questions are defined in this SLR study:

1. What are the key motivations for involving fairness in artificial intelligence?
2. Which fairness definitions have been taken into account for fair artificial intelligence?
3. Which methodologies have been proposed as a fair artificial intelligence model?
4. What are the drawbacks, challenges, and possible solutions in fair artificial intelligence?

The contributions of this study are as follows:

1. We have evaluated 85 research papers from different dimensions and responded using different categories for each research question.
2. Challenges and possible solutions have been also discussed in this work; this might pave the way for further research.

**2.2.1 Research methodology.** We conducted the SLR following Kitchenham's Evidence-Based Software Engineering guideline for software engineering (Dybå, Kitchenham, & Jorgensen, 2005). The guideline has been adapted to reflect the specific problems of software engineering research and covers three phases of a systematic review; planning the review; conducting the review, and reporting the review. Figure 3 illustrates our systematical literature review with divided sections.

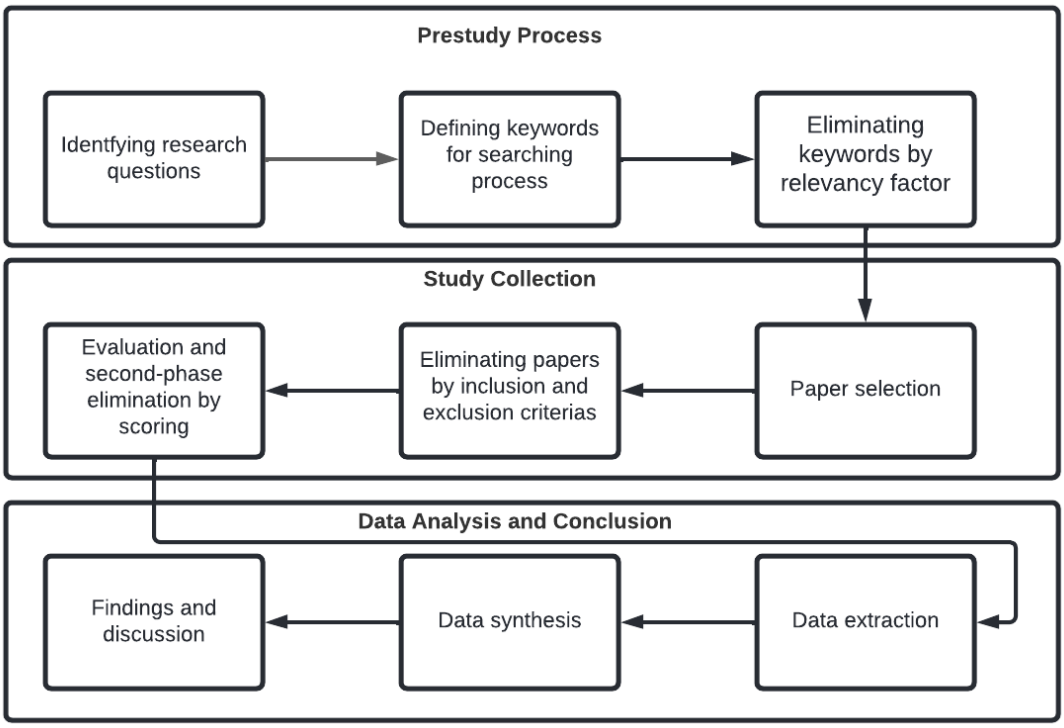


*Figure 3.* Systematic review process

**2.2.2 Research questions.** This research goal is to analyze published studies and their findings on the fair artificial intelligence term and methodologies. Table 5 exhibits the research questions we have found to make our systematic literature review paper more focused.

Table 5

*Research Questions*

| ID | Research Question |
|---|---|
| RQ1 | What are the areas where fair artificial intelligence has been used? |
| RQ2 | Which fairness definitions have been taken into account for fair artificial intelligence? |
| RQ3 | Which methodologies have been proposed as a fair artificial intelligence model? |
| RQ4 | What are the drawbacks, challenges, and possible solutions in fair artificial intelligence? |

**2.2.3 Search strategy.** After defining the research question, we have created a word list that can be useful for the string search from most related to less. Alternative terms are connected using OR operator to obtain a wider range of results. There are four search segments for the search strategy (i) creating a key union about artificial intelligence; (ii) creating a key union for fairness; (iii) creating a key union for fairness in artificial intelligence systems; and (iv) and state-of-the-art works about fair artificial intelligence. With derivatives and intersections of four segments, we have obtained keyword- space which can be seen below.

Keyword-Space: machine learning, artificial intelligence, fairness, algorithmic fairness, group fairness, individual fairness, demographic parity, equalized opportunity, equalized odds groups, biased artificial intelligence models, gender-biased artificial intelligence models, algorithmic fairness, mitigating bias in artificial intelligence models, unfair artificial intelligence, fair artificial intelligence models

**2.2.4 Literature resources.** For resource collection, several sources were used:

Paper databases:

- ScienceDirect: https://www.sciencedirect.com/
- IEEE Xplore: https://ieeexplore.ieee.org/

Search engines

- Google Scholar: https://scholar.google.com/
- Microsoft Academic: https://academic.microsoft.com/
- WorldWideScience: https://worldwidescience.org/

Due to the increasing popularity of artificial intelligence, millions of articles have been published in this area therefore we had to limit our keywords to ("Artificial Intelligence") AND (“Fairness” OR “Mitigating Bias” OR "Unbiased OR "Justice" OR "Fair").

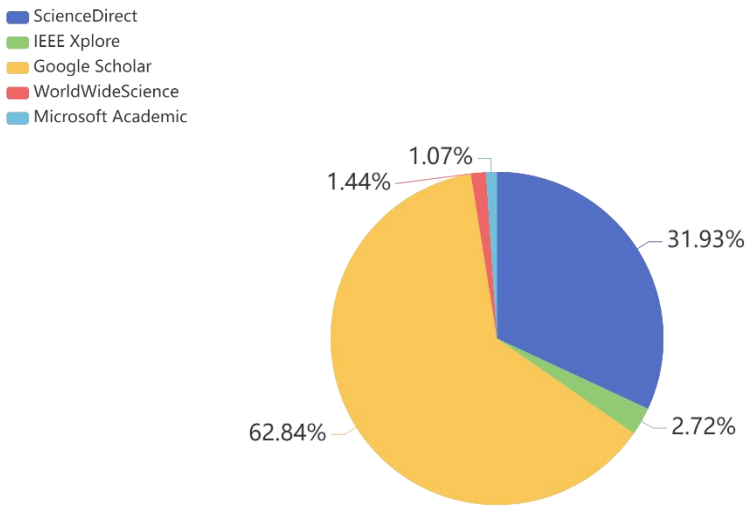


*Figure 4.* Distribution of collecting papers by source after first phase

With this search we have obtained 66534 paper results, Figure 4 depicts that an overwhelming number of 41700 articles have been found in Google Scholar, while other sources Google Scholar, Science Direct, IEEE Xplore, WorldWideScience, and Microsoft Academic got returned 21186, 1802, 958, and 708 paper results respectively.

To eliminate the number of irrelevant or non-convenient papers, we have applied the filtering process with criteria that have been determined by authors and have been applied incrementally. After this elimination, Figure 5 shows that results have been decreased to 3908 papers where the results found in Google Scholar have been reduced to 1520, papers in IEEE Xplore to 1256, papers in Microsoft Academic to 536, and papers in ScienceDirect to 515 and WorldWideScience to 81. By using article search engines, we can scan different databases we have taken into account such as IEEE articles, and Science Direct journals, and not taken into account such as ResearchGate, Wiley, Springer journals, and so on. Therefore, we will present the total number of articles to avoid confusion in publication numbers due to duplicated papers in different sources, instead of categorizing counts by sources.

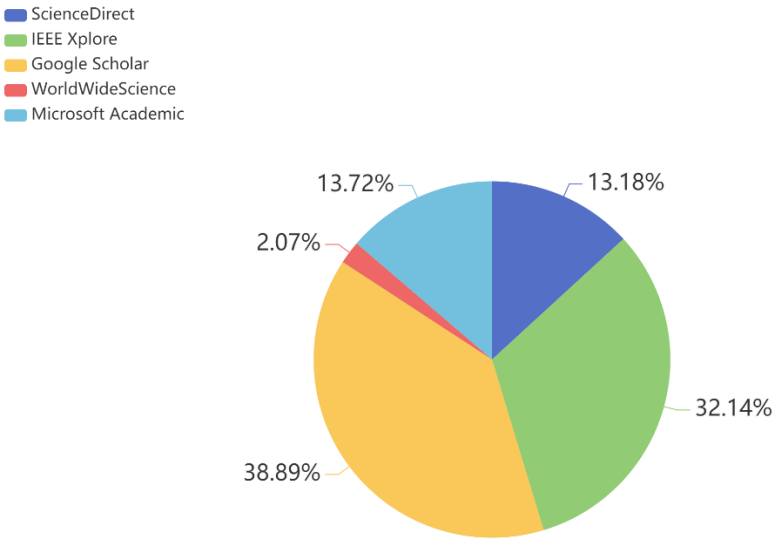


*Figure 5.* Distribution of collecting papers by source after second phase

We eliminated all review articles, correspondence articles, discussion papers, and duplicated papers from different sources. After these criteria, the total number of articles was reduced to 120. Table 6 and Table 7 illustrate the inclusion and exclusion criteria for filtering papers.

Table 6

*Inclusion Criteria*

| ID | Criterion |
|---|---|
| INCR1 | The publication includes fair machine learning methodology |
| INCR2 | The publication is a primary study |

Table 7

*Ex-clusion Criteria*

| ID | Criterion |
|---|---|
| EXCR1 | Not related to both fairness and artificial intelligence |
| EXCR2 | Not published in English |

Table 7 (continued)

| ID | Criterion |
|---|---|
| EXCR3 | A survey or a review publication |
| EXCR4 | Duplicated publication |
| EXCR5 | Archive papers which are non-peer-reviewed |
| EXCR6 | The publication is older than 2018 |

After the elimination phase, selected articles have been subjected to quality testing to make sure that only high-quality publications are being used. Eight assessment questions were used from the study of (Slob, Catal, & Kassahun, 2021), where this set of questions is widely used in SLR papers. Table 8 shows the assessment questions for the assessment of the selected papers.

Table 8

*Quality Assessment Questions*

| ID | Assessments questions |
|---|---|
| QAQ1 | Are the aims of the study clearly stated? |
| QAQ2 | Are the scope and context of the study clearly defined? |
| QAQ3 | Is the proposed solution clearly explained and validated by an empirical study? |
| QAQ4 | Are the variables used in the study likely to be valid and reliable? |
| QAQ5 | Is the research process documented adequately? |
| QAQ6 | Are all study questions answered? |
| QAQ7 | Are the negative findings presented? |
| QAQ8 | Are the main findings stated clearly in terms of creditability, validity, and reliability? |

Each question is assessed with three-point scale of 1 (yes), 0 (no), 0.5 (partial). Therefore, 0 is the minimum score, and 8 is the maximum score for a paper. A paper with a total score of 4 or lower was excluded. Table 9 includes our question-based assessment for each article.

Table 9

*Detailed Assessment Table*

| Article | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Overall |
|---|---|---|---|---|---|---|---|---|---|
| FairNN - Conjoint Learning… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,0 | 1,0 | 7,0 |
| Learning Unbiased Represen… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 8,0 |
| FARF: A Fair and Adaptive… | 0,5 | 0,5 | 1,0 | 1,0 | 1,0 | 0,5 | 0,0 | 0,5 | 5,0 |
| Fairness in Network Represe… | 1,0 | 1,0 | 0,5 | 0,5 | 0,5 | 1,0 | 0,5 | 1,0 | 6,0 |
| *Evaluating new technology f…* | 1,0 | 1,0 | 0,0 | 0,5 | 0,5 | 0,5 | 0,0 | 0,5 | 4,0 |
| Toward Learning Trustworth… | 0,5 | 0,5 | 1,0 | 1,0 | 1,0 | 1,0 | 0,0 | 1,0 | 6,0 |
| *From Learning to Relearning…* | 0,5 | 0,5 | 0,0 | 0,5 | 0,0 | 1,0 | 0,5 | 0,0 | 3,0 |
| Low-Shot Deep Learning of… | 1,0 | 1,0 | 0,5 | 0,5 | 0,5 | 0,5 | 1,0 | 1,0 | 6,0 |
| *Towards Formal Fairness in…* | 0,5 | 0,5 | 0,0 | 0,5 | 0,5 | 0,5 | 0,0 | 0,0 | 2,5 |
| FairFaceGAN: Fairness-awa… | 0,5 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,0 | 1,0 | 6,5 |
| Predict Responsibly: Improv… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 1,0 | 7,5 |
| *Fairness in Supervised Learn…* | 1,0 | 1,0 | 0,0 | 0,0 | 0,5 | 0,0 | 0,0 | 0,0 | 2,5 |
| *Path-Specific Counterfactual…* | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,0 | 0,5 | 3,5 |
| *Pooling of Causal Models u…* | 0,5 | 0,5 | 0,5 | 0,5 | 1,0 | 0,5 | 0,5 | 0,0 | 4,0 |
| Fairness GAN | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,0 | 1,0 | 7,0 |
| FairGAN: Fairness-aware… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 1,0 | 7,5 |
| Multiaccuracy: Black-Box… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 1,0 | 7,5 |
| Fairness through Causal… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 8,0 |
| *Using Image Fairness…* | 1,0 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,0 | 0,5 | 4,0 |
| Active Fairness in Algorithm… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,0 | 1,0 | 7,0 |
| Policy Learning for Fairness… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,0 | 0,5 | 6,0 |
| Mathematical Notions vs…. | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 8,0 |
| *Fairness in Recommendation…* | 0,5 | 1,0 | 0,5 | 0,5 | 0,5 | 0,5 | 0,0 | 0,5 | 4,0 |
| Contrastive Fairness in… | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,5 | 0,5 | 0,5 | 6,0 |
| CERTIFAI: Counterfactual… | 1,0 | 1,0 | 1,0 | 0,5 | 1,0 | 1,0 | 0,0 | 1,0 | 6,5 |
| Multi-Armed Bandits with… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 8,0 |
| FairNAS: Rethinking Evalu… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,0 | 1,0 | 7,0 |
| Fairness-enhancing interve… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,0 | 0,5 | 6,0 |
| FAHT: An Adaptive Fairnes… | 0,5 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,0 | 0,5 | 5,5 |

Table 9 (continued)

| Article | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Overall |
|---|---|---|---|---|---|---|---|---|---|
| *Approaching Machine Learn…* | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,0 | 0,5 | 3,5 |
| The Impact of Data Preparati… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,5 | 1,0 | 7,0 |
| Fairness Sample Complexity… | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,5 | 0,0 | 1,0 | 6,0 |
| *Toward a better trade-off…* | 1,0 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,0 | 0,5 | 4,0 |
| Fairness Violations and… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,5 | 7,0 |
| Maintaining Discrimination… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,0 | 0,5 | 6,5 |
| Towards Fairness in Visual… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 8,0 |
| Leveraging Semi-Supervised… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,0 | 0,5 | 6,5 |
| FAE: A Fairness-Aware… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,0 | 0,5 | 6,0 |
| Learning Fairness-aware… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,0 | 1,0 | 7,0 |
| Fairness by Learning Orthogo... | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 1,0 | 7,5 |
| FairMOT: On the Fairness of… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,0 | 1,0 | 7,0 |
| Deepfair: Deep Lear… | 0,5 | 0,5 | 0,5 | 1,0 | 1,0 | 1,0 | 0,0 | 0,5 | 5,0 |
| Achieving Fairness via Post… | 0,5 | 1,0 | 0,5 | 0,5 | 0,5 | 0,5 | 0,0 | 0,5 | 4,0 |
| Fairness With Overlapping… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,0 | 1,0 | 7,0 |
| Ensuring Fairness Beyond the.. | 1,0 | 1,0 | 0,5 | 1,0 | 1,0 | 0,5 | 0,0 | 0,5 | 5,5 |
| Investigating Bias and Fairne | 1,0 | 1,0 | 0,5 | 1,0 | 1,0 | 0,5 | 0,5 | 0,5 | 6,0 |
| Accuracy and Fairness Trad… | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,0 | 0,5 | 3,5 |
| Collaborative Fairness in… | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,5 | 0,0 | 0,5 | 5,5 |
| Fair and accurate age predi… | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,5 | 0,5 | 1,0 | 6,5 |
| Group Fairness by Probabil… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,0 | 0,5 | 6,0 |
| Fairness in Semi-supervised… | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 1,0 | 0,0 | 1,0 | 6,5 |
| Differentially Private… | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 1,0 | 0,0 | 1,0 | 6,5 |
| *Bridging Machine Learning…* | 1,0 | 0,0 | 0,5 | 1,0 | 0,5 | 0,0 | 0,0 | 0,5 | 3,5 |
| Augmented Fairness: An… | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 1,0 | 0,0 | 1,0 | 6,5 |
| Exploring Text Specific and… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 8,0 |
| FairOD: Fairness-aware… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,0 | 1,0 | 7,0 |
| *Improving the Fairness of…* | 1,0 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,0 | 0,5 | 4,0 |
| TARA: Training and… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,0 | 1,0 | 6,5 |

Table 9 (continued)

| Article | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Overall |
|---|---|---|---|---|---|---|---|---|---|
| *dalex: Responsible Machine…* | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,0 | 0,5 | 3,5 |
| *Fairness in Machine Learning* | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,0 | 0,0 | 1,0 | 3,5 |
| *Fairness in Cardiac Magne…* | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 4,0 |
| *Fairness-Aware PAC Learni…* | 1,0 | 0,5 | 0,5 | 0,0 | 0,5 | 0,5 | 0,0 | 1,0 | 4,0 |
| Evaluating Fairness of Mach… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,0 | 1,0 | 7,0 |
| *Fairness In Tabnet Model By…* | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 1,0 | 1,0 |
| *Estimating and Improving…* | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,0 | 1,0 | 4,0 |
| *Tilted Cross-Entropy (TCE)…* | 1,0 | 1,0 | 1,0 | 0,5 | 0,5 | 0,0 | 0,0 | 0,0 | 4,0 |
| *AI Fairness via Domain…* | 1,0 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,0 | 0,0 | 3,5 |
| *Can Active Learning…* | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,0 | 0,0 | 3,0 |
| Biased Edge Dropout for… | 1,0 | 0,5 | 0,5 | 1,0 | 1,0 | 1,0 | 0,5 | 1,0 | 6,5 |
| *Information Theoretic…* | 1,0 | 0,5 | 0,5 | 0,5 | 0,5 | 1,0 | 0,0 | 0,0 | 4,0 |
| Fairness-Aware Unsupervise… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,0 | 1,0 | 6,5 |
| Understanding and Improvi… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,0 | 1,0 | 7,0 |
| Consistent Instance False… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 1,0 | 7,5 |
| Fairness in Cardiac MR Ima… | 1,0 | 1,0 | 0,5 | 0,5 | 1,0 | 1,0 | 0,0 | 1,0 | 6,0 |
| Balancing Accuracy and… | 1,0 | 1,0 | 0,5 | 0,5 | 1,0 | 1,0 | 0,0 | 1,0 | 6,0 |
| *The Sharpe predictor for…* | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,0 | 0,5 | 3,5 |
| Fairness-Aware Online Met… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,0 | 1,0 | 6,5 |
| Style Pooling: Automatic Te… | 1,0 | 1,0 | 0,5 | 0,5 | 0,5 | 1,0 | 0,5 | 1,0 | 6,0 |
| *A Fairness Analysis on Priv…* | 1,0 | 1,0 | 0,5 | 0,5 | 0,5 | 0,0 | 0,0 | 0,5 | 4,0 |
| Fairness-aware Class… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,0 | 0,5 | 6,5 |
| Enhancing Model Robustnes… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,0 | 0,5 | 6,0 |
| xFAIR: Better Fairness via… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,0 | 0,5 | 6,5 |
| *A Burden Shared is a Burde…* | 1,0 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,0 | 0,5 | 4,0 |
| One-Network Adversarial… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,0 | 0,5 | 6,0 |
| Path-Specific Counterfactual… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,0 | 0,5 | 6,0 |
| Non-Discriminatory Mach… | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,5 | 0,0 | 0,5 | 5,5 |
| Learning Disentangled… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,0 | 1,0 | 7,0 |

Table 9 (continued)

| Article | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Overall |
|---|---|---|---|---|---|---|---|---|---|
| A Cluster-based Solution to… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,0 | 0,5 | 6,0 |
| Adaptive Sensitive Reweight… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,0 | 0,5 | 6,0 |
| *Fairness of Extractive Text…* | 0,5 | 0,5 | 0,0 | 0,5 | 0,5 | 0,5 | 0,0 | 0,5 | 3,0 |
| Fairness-Aware Tensor-Base… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,0 | 1,0 | 7,0 |
| Fairness Warnings and Fair… | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,5 | 0,0 | 1,0 | 6,0 |
| *Gender Slopes: Counterfactu…* | 1,0 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,0 | 0,5 | 4,0 |
| Achieving Causal Fairness… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,0 | 1,0 | 6,5 |
| *Improving Prediction Fairne…* | 1,0 | 0,5 | 0,5 | 0,0 | 0,5 | 1,0 | 0,0 | 0,5 | 4,0 |
| *Towards Formal Fairness in…* | 1,0 | 1,0 | 0,5 | 0,5 | 0,5 | 0,0 | 0,0 | 0,5 | 4,0 |
| *Context-conscious fairness…* | 1,0 | 1,0 | 0,5 | 0,5 | 0,5 | 0,0 | 0,0 | 0,5 | 4,0 |
| Local Data Debiasing for… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 1,0 | 7,5 |
| Fair navigation planning: a… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 1,0 | 7,5 |
| Fairness-Aware Training of… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,0 | 0,5 | 6,0 |
| Fairness and Bias in Onli… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,0 | 0,5 | 6,0 |
| Does Robustness Improve… | 1,0 | 1,0 | 0,5 | 0,5 | 0,5 | 0,5 | 0,0 | 0,5 | 4,5 |
| FEAT: A Fairness-enhancin… | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,5 | 0,0 | 0,5 | 5,5 |
| Interventional Fairness with… | 1,0 | 1,0 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,0 | 4,5 |
| FABBOO - Online Fairnes… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,0 | 0,5 | 6,5 |
| Fairbatch: Batch Selection… | 1,0 | 1,0 | 1,0 | 0,5 | 0,5 | 0,5 | 0,0 | 0,5 | 5,0 |
| Fairness via Representation… | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 0,5 | 0,0 | 1,0 | 6,0 |
| Gradient-Driven Rewards to… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,0 | 0,5 | 6,5 |
| Exploiting Transferable… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,0 | 0,5 | 6,5 |
| Joint Transfer of Model… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,0 | 1,0 | 7,0 |
| Individual Fairness for Grap… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,0 | 1,0 | 7,0 |
| Enhancing Long Term… | 1,0 | 1,0 | 1,0 | 1,0 | 0,5 | 1,0 | 0,0 | 1,0 | 6,5 |
| *The Cost of Fairness in Bina…* | 1,0 | 1,0 | 0,5 | 0,5 | 0,0 | 0,0 | 0,0 | 0,5 | 3,5 |
| Neural-Symbolic Integratio… | 1,0 | 1,0 | 0,5 | 0,5 | 0,0 | 0,5 | 0,5 | 1,0 | 5,0 |
| Online Set Selection with… | 1,0 | 1,0 | 1,0 | 0,5 | 0,5 | 0,5 | 0,0 | 1,0 | 5,5 |
| Imparting Fairness to Pre… | 1,0 | 1,0 | 1,0 | 0,5 | 0,5 | 1,0 | 0,0 | 1,0 | 6,0 |

Table 9 (continued)

| Article | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Overall |
|---|---|---|---|---|---|---|---|---|---|
| *Addressing Fairness in…* | 0,5 | 0,5 | 0,5 | 0,0 | 0,0 | 0,5 | 0,0 | 0,5 | 2,5 |
| Subgroup Generalization and… | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,0 | 0,5 | 6,5 |
| *The accuracy, fairness, and…* | 0,5 | 0,5 | 0,5 | 0,0 | 0,0 | 0,5 | 0,0 | 0,5 | 2,5 |
| Achieving Differential… | 1,0 | 1,0 | 0,5 | 0,5 | 0,5 | 0,5 | 0,0 | 1,0 | 5,0 |

The studies shown in italics in the Table 9 are the studies that were eliminated as a result of the evaluation.

We provide graphical representation to present the statistical implications of article scores. Figure 6 depicts the assessment point distribution for selected papers, and 85 of them are good candidates for our systematic review. Therefore, deductions and conclusions presented in this systematic literature review are based on 85 articles.

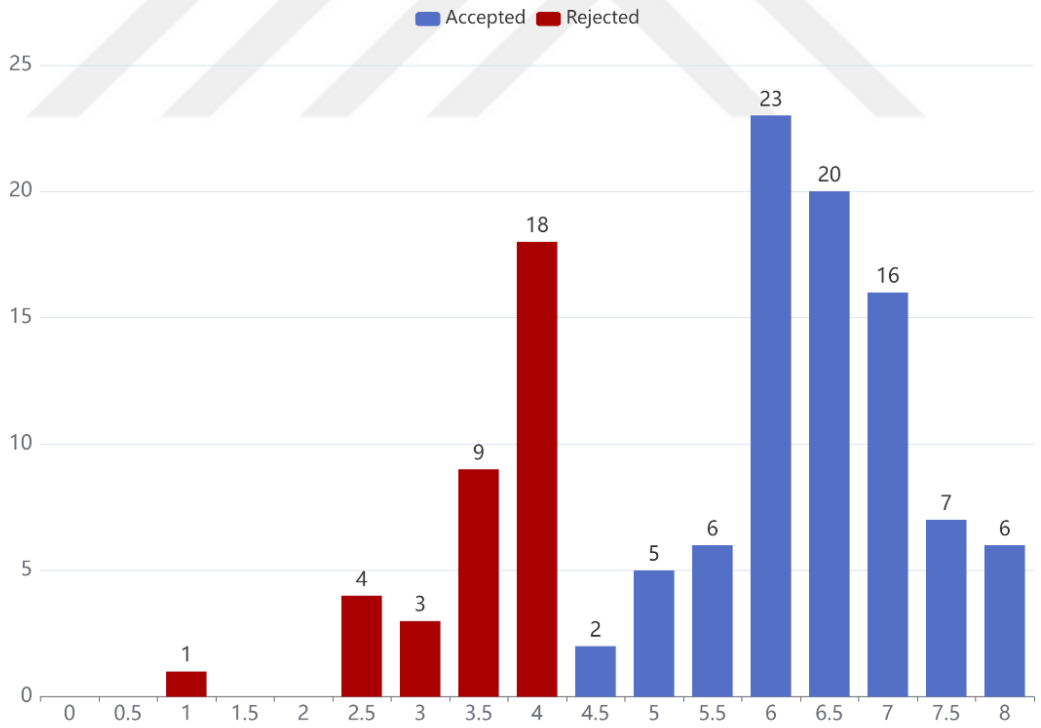


*Figure 6.* Distribution of papers by assessment score

**2.2.5 Data extraction.** After selecting suitable candidates, data relevant to the research questions were extracted, stored, and categorized in a spreadsheet that contains additional information about the papers. Papers were read in full

comparatively, and required data was collected. The collection has been categorized into different groups based on questions. The findings of the questions to be explained in detail can be summarized as follows. In RQ1, motivations were categorized as: bias in decision-making systems, computer vision, medicine, natural language processing, and recommendation systems. The most commonly used measures of fairness and metrics in RQ2 are listed as follows: demographic parity, equalized odds, equal opportunity, statistical parity, and individual fairness. The machine learning models used in RQ3 to realize fair AI or test the proposed system, in order of percentage usage, can be categorized as decision trees, logistic regression models, generative adversarial networks, ResNet architectures, and auto-encoders. The problems and solutions in the last question, namely RQ4, can be summarized as follows: the challenges can be categorized as extendability, interaction, real-world differences, machine learning process, the trade-off between fairness and accuracy, and domain adaptation. In addition to the answers to these questions, information including general information about the article was also collected and the format of the collected information is presented in Table 10. Table 11 contains details of the extracted information for each article without questions. We will present questions answers of each article in Table 12.

Table 10

*Data Extraction Format*

| No | Extraction elements |
|---|---|
| 1 | ID |
| 2 | Title |
| 3 | Link |
| 4 | Year |
| 5 | Publisher |
| 6 | Publication channel |
| 7 | Conference |
| 8 | Domains |
| 9 | Fairness measurements |

Table 10 (continued)

| No | Extraction elements |
|---|---|
| 10 | Fair machine learning methodologies |
| 11 | Challenges and possible solutions |

Table 11

*Details of Selected Publications*

| ID | Title | Year | Publisher | Author |
|---|---|---|---|---|
| 1 | FairNN - Conjoint Learning of Fair Representations for Fair Decisions | 2020 | Springer International Publishing | Tongxin Hu, Vasileios Iosifidis, Wentong Liao, Hang Zhang, Michael YingYang, Eirini Ntoutsi, and Bodo Rosenhahn |
| 2 | Learning Unbiased Representations via Rényi Minimization | 2021 | Springer International Publishing | Vincent Grari, Oualid El Hajouji, Sylvain Lamprier, Marcin Detyniecki |
| 3 | FARF: A Fair and Adaptive Random Forests Classifier | 2021 | Springer International Publishing | Wenbin Zhang, Albert Bifet, Xiangliang Zhang, Jeremy C. Weiss, Wolfgang Nejdl |
| 4 | Fairness in Network Representation by Latent Structural Heterogeneity in Observational Data | 2020 | AAAI Press | Xin Du, Yulong Pei, Wouter Duivesteijn, Mykola Pechenizkiy |
| 5 | Toward Learning Trustworthily from Data Combining Privacy, Fairness, and Explainability: An Application to Face Recognition | 2021 | MDPI | Danilo Franco, Luca Oneto, Nicolò Navarin, Davide Anguita |

Table 11 (continued)

| ID | Title | Year | Publisher | Author |
|---|---|---|---|---|
| 6 | Low-Shot Deep Learning of Diabetic Retinopathy With Potential Applications to Address Artificial Intelligence Bias in Retinal Diagnostics and Rare Ophthalmic Diseases | 2020 | American Medical Association | Philippe Burlina, William Paul, Philip Mathew; Neil Joshi, Katia D. Pacheco, Neil M. Bressler |
| 7 | FairFaceGAN: Fairness-aware Facial Image-to-Image Translation | 2020 | BMVA Press | Sunhee Hwang, Sungho Park, Dohyung Kim, Mirae Do, Hyeran Byun |
| 8 | Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer | 2018 | Curran Associates, Inc. | David Madras, Toniann Pitassi, Richard Zemel |
| 9 | Fairness GAN | 2019 | IBM | Prasanna Sattigeri, Samuel C. Hoffman, Vijil Chenthamarakshan, Kush R. Varshney |
| 10 | FairGAN: Fairness-aware Generative Adversarial Networks | 2020 | Springer International Publishing | Tongxin Hu, Vasileios Iosifidis, Wentong Liao, Hang Zhang, Michael YingYang, Eirini Ntoutsi, and Bodo Rosenhahn |
| 11 | Multiaccuracy: Black-Box Post-Processing for Fairness in Classification | 2018 | Curran Associates, Inc. | Depeng Xu, Shuhan Yuan, Lu Zhang, Xintao Wu |
| 12 | Fairness through Causal Awareness: Learning Causal Latent-Variable Models for Biased Data | 2019 | Association for Computing Machinery | Michael P. Kim, Amirata Ghorbani, James Zou |

Table 11 (continued)

| ID | Title | Year | Publisher | Author |
|---|---|---|---|---|
| 13 | Active Fairness in Algorithmic Decision Making | 2019 | Association for Computing Machinery | David Madras, Elliot Creager, Toniann Pitassi, Richard Zemel |
| 14 | Policy Learning for Fairness in Ranking | 2019 | Association for Computing Machinery | Alejandro Noriega-Campero, Michiel A. Bakker, Bernardo Garcia-Bulle, Alex Pentland |
| 15 | Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning | 2018 | Curran Associates, Inc. | Ashudeep Singh, Thorsten Joachims |
| 16 | Contrastive Fairness in Machine Learning | 2019 | Association for Computing Machinery | Megha Srivastava, Hoda Heidari, Andreas Krause |
| 17 | CERTIFAI: Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence models | 2020 | IEEE | Tapabrata Chakraborti, Arijit Patra, Alison Noble |
| 18 | Multi-Armed Bandits with Fairness Constraints for Distributing Resources to Human Teammates | 2020 | Association for Computing Machinery | Shubham Sharma, Jette Henderson, Joydeep Ghosh |
| 19 | FairNAS: Rethinking Evaluation Fairness of Weight Sharing NeuralArchitecture Search | 2020 | Association for Computing Machinery | Houston Claure, Yifang Chen, Jignesh Modi, Malte Jung, Stefanos Nikolaidis |
| 20 | Fairness-enhancing interventions in stream classification | 2019 | IEEE | Xiangxiang Chu, Bo Zhang, Ruijun Xu |

Table 11 (continued)

| ID | Title | Year | Publisher | Author |
|---|---|---|---|---|
| 21 | FAHT: An Adaptive Fairness-aware Decision Tree Classifier | 2019 | IJCAI | Wenbin Zhang, Eirini Ntoutsi |
| 22 | The Impact of Data Preparation on the Fairness of Software Systems | 2019 | Curran Associates, Inc. | Ines Valentim, Nuno Lourenc¸o, Nuno Antunes |
| 23 | Fairness Sample Complexity and the Case for Human Intervention | 2019 | Not Specified | Ananth Balashankar, Alyssa Lees |
| 24 | Fairness Violations and Mitigation under Covariate Shift | 2021 | Association for Computing Machinery | Harvineet Singh, Rina Singh, Vishwali Mhasawade, Rumi Chunara |
| 25 | Maintaining Discrimination and Fairness in Class Incremental Learning | 2020 | Curran Associates, Inc. | Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, Shutao Xia |
| 26 | Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation | 2020 | Curran Associates, Inc. | Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, Olga Russakovsky |
| 27 | Leveraging Semi-Supervised Learning for Fairness using Neural Networks | 2019 | Curran Associates, Inc. | Vahid Noroozi, Sara Bahaadini, Samira Sheikhi, Nooshin Mojab, Philip S. Yu |
| 28 | FAE: A Fairness-Aware Ensemble Framework | 2019 | Curran Associates, Inc. | Vasileios Iosifidis, Besnik Fetahu, Eirini Ntoutsi |
| 29 | Learning Fairness-aware Relational Structures | 2020 | IOS Press | Yue Zhang, Arti Ramesh |
| 30 | Fairness by Learning Orthogonal Disentangled Representations | 2020 | Springer International Publishing | Mhd Hasan Sarhan, Nassir Navab, Abouzar Eslami, Shadi Albarqouni |

Table 11 (continued)

| ID | Title | Year | Publisher | Author |
|---|---|---|---|---|
| 31 | FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking | 2020 | Springer International Publishing | Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, Wenyu Liu |
| 32 | Deepfair: Deep Learning For Improving Fairness In Recommender Systems | 2020 | Universidad Internacional de La Rioja | Jesús Bobadilla, Raúl Lara-Cabrera, Ángel González-Prieto, Fernando Ortega |
| 33 | Fairness With Overlapping Groups | 2020 | Not Specified | Forest Yang, Moustapha Cisse, Sanmi Koyejo |
| 34 | Ensuring Fairness Beyond the Training Data | 2020 | Not Specified | Debmalya Mandal, Samuel Deng, Suman Jana, Jeannette M. Wing, Daniel Hsu |
| 35 | Investigating Bias and Fairness in Facial Expression Recognition | 2020 | Springer International Publishing | Tian Xu, Jennifer White, Sinan Kalkan, Hatice Gunes |
| 36 | Collaborative Fairness in Federated Learning | 2020 | Springer International Publishing | Lingjuan Ly, Xinyi Xu, Qian Wang |
| 37 | Fair and accurate age prediction using distribution aware data curation and augmentation | 2021 | Not Specified | Yushi Cao, David Berend, Palina Tolmach, Guy Amit, Moshe Levy, Yang Liu, Asaf Shabtai, Yuval Elovici |
| 38 | Group Fairness by Probabilistic Modeling with Latent Fair Decisions | 2021 | AAAI Press | YooJung Choi, Meihua Dang, Guy Van den Broeck |
| 39 | Fairness in Semi-supervised Learning: Unlabeled Data Help to Reduce Discrimination | 2020 | IEEE | Tao Zhang, Tianqing Zhu, Jing Li, Mengde Han, Wanlei Zhou, Philip S. Yu |

Table 11 (continued)

| ID | Title | Year | Publisher | Author |
|---|---|---|---|---|
| 40 | Differentially Private Representation for NLP: Formal Guarantee and An Empirical Study on Privacy and Fairness | 2020 | Association for Computational Linguistics | Lingjuan Lyu, Xuanli He, Yitong Li |
| 41 | Augmented Fairness: An Interpretable Model Augmenting Decision-Makers' Fairness | 2020 | Not Specified | Tong Wang, Maytal Saar-Tsechansky |
| 42 | Exploring Text Specific and Blackbox Fairness Algorithms in Multimodal Clinical NLP | 2020 | Association for Computational Linguistics | John Chen, Ian Berlot-Attwell, Safwan Hossain, Xindi Wang, Frank Rudzicz |
| 43 | FairOD: Fairness-aware Outlier Detection | 2021 | Association for Computing Machinery | Shubhranshu Shekhar, Neil Shah, Leman Akoglu |
| 44 | TARA: Training and Representation Alteration for AI Fairness and Domain Generalization | 2022 | Journals Gateway | William Paul, Armin Hadzic, Neil Joshi, Fady Alajaji, Phil Burlina |
| 45 | Evaluating Fairness of Machine Learning Models Under Uncertain and Incomplete Information | 2021 | Association for Computing Machinery | Pranjal Awasthi, Alex Beutel, Matthaeus Kleindessner, Jamie Morgenstern, Xuezhi Wang |
| 46 | Biased Edge Dropout for Enhancing Fairness in Graph Representation Learning | 2021 | IEEE | Indro Spinelli, Simone Scardapane, Amir Hussain, Aurelio Uncini |

Table 11 (continued)

| ID | Title | Year | Publisher | Author |
|---|---|---|---|---|
| 47 | Fairness-Aware Unsupervised Feature Selection | 2021 | Association for Computing Machinery | Xiaoying Xing, Hongfu Liu, Chen Chen, Jundong Li |
| 48 | Understanding and Improving Fairness-Accuracy Trade-offs in Multi-Task Learning | 2021 | Association for Computing Machinery | Yuyan Wang, Xuezhi Wang, Alex Beutel, Flavien Prost, Jilin Chen, Ed H. Chi |
| 49 | Consistent Instance False Positive Improves Fairness in Face Recognition | 2021 | IEEE Computer Society | Xingkun Xu, Yuge Huang, Pengcheng Shen, Shaoxin Li, Jilin Li, Feiyue Huang, Yong Li, Zhen Cui |
| 50 | Fairness in Cardiac MR Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation | 2021 | Springer International Publishing | Esther Puyol-Anton, Bram Ruijsink, Stefan K. Piechnik, Stefan Neubauer, Steffen E. Petersen, Reza Razavi, Andrew P. King |
| 51 | Balancing Accuracy and Fairness for Interactive Recommendation with Reinforcement Learning | 2020 | Springer International Publishing | Weiwen Liu, Feng Liu, Ruiming Tang, Ben Liao, Guangyong Chen, Pheng Ann Heng |
| 52 | Fairness-Aware Online Meta-learning | 2021 | Association for Computing Machinery | Chen Zhao, Feng Chen, Bhavani Thuraisingham |
| 53 | Style Pooling: Automatic Text Style Obfuscation for Improved Classification Fairness | 2021 | Association for Computational Linguistics | Fatemehsadat Mireshghallah, Taylor Berg-Kirkpatrick |
| 54 | Fairness-aware Class Imbalanced Learning | 2021 | Association for Computational Linguistics | Shivashankar Subramanian, Afshin Rahimi, Timothy Baldwin, Trevor Cohn, Lea Frermann |

Table 11 (continued)

| ID | Title | Year | Publisher | Author |
|---|---|---|---|---|
| 55 | Enhancing Model Robustness and Fairness with Causality: A Regularization Approach | 2021 | Association for Computational Linguistics | Zhao Wang, Kai Shu, Aron Culotta |
| 56 | Fairness Degrading Adversarial Attacks Against Clustering Algorithms | 2021 | Not Specified | Anshuman Chhabra, Adish Singla, Prasant Mohapatra |
| 57 | One-Network Adversarial Fairness | 2019 | AAAI Press | Tameem Adel, Isabel Valera, Zoubin Ghahramani, Adrian Weller |
| 58 | Path-Specific Counterfactual Fairness | 2019 | AAAI Press | Silvia Chiappa, Thomas P. S. Gillam |
| 59 | Non-Discriminatory Machine Learning Through Convex Fairness Criteria | 2018 | Association for Computing Machinery | Naman Goel, Mohammad Yaghini, Boi Faltings |
| 60 | Learning Disentangled Representation for Fair Facial Attribute Classification via Fairness-aware Information Alignment | 2021 | AAAI Press | Sungho Park, Sunhee Hwang, Dohyung Kim, Hyeran Byun |
| 61 | A Cluster-based Solution to Achieve Fairness in Federated Learning | 2020 | IEEE | Fengpan Zhao, Yan Huang, Akshita Maradapu Vera Venkata Sai, Yubao Wu |
| 62 | Adaptive Sensitive Reweighting to Mitigate Bias in Fairness-aware Classification | 2018 | International World Wide Web Conferences Steering Committee | Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, Yiannis Kompatsiaris |

Table 11 (continued)

| ID | Title | Year | Publisher | Author |
|---|---|---|---|---|
| 63 | Fairness-Aware Tensor-Based Recommendation | 2018 | Association for Computing Machinery | Ziwei Zhu, Xia Hu, James Caverlee |
| 64 | Fairness Warnings and Fair-MAML: Learning Fairly with Minimal Data | 2020 | Association for Computing Machinery | Dylan Slack, Sorelle Friedler, Emile Givental |
| 65 | Achieving Causal Fairness through Generative Adversarial Networks | 2019 | AAAI Press | Depeng Xu, Yongkai Wu, Shuhan Yuan, Lu Zhang, Xintao Wu |
| 66 | Local Data Debiasing for Fairness Based on Generative Adversarial Training | 2021 | MDPI | Ulrich Aïvodji , François Bidet , Sébastien Gambs , Rosin Claude Ngueveu, Alain Tapp |
| 67 | Fair navigation planning: a resource for characterizing and designing fairness in mobile robots | 2020 | Elsevier | Martim Brandão, Marina Jirotka, Helena Webb, Paul Luff |
| 68 | Fairness-Aware Training of Decision Trees by Abstract Interpretation | 2021 | Association for Computing Machinery | Francesco Ranzato, Caterina Urban, Marco Zanella |
| 69 | Fairness and Bias in Online Selection | 2021 | PMLR | Jose Correa, Andres Cristi, Paul Duetting, Ashkan Norouzi-Fard |
| 70 | Does Robustness Improve Fairness? Approaching Fairness with Word Substitution Robustness Methods for Text Classification | 2021 | Association for Computational Linguistics | Yada Pruksachatkun, Satyapriya Krishna, Jwala Dhamala, Rahul Gupta, Kai-Wei Chang |

Table 11 (continued)

| ID | Title | Year | Publisher | Author |
|---|---|---|---|---|
| 71 | FEAT: A Fairness-Enhancing and Concept-Adapting Decision Tree Classifier | 2020 | Springer International Publishing | Wenbin Zhang, Albert Bifet |
| 72 | Interventional Fairness with Indirect Knowledge of Unobserved Protected Attributes | 2021 | MDPI | Sainyam Galhotra, Karthikeyan Shanmugam, Prasanna Sattigeri, Kush R Varshne |
| 73 | FABBOO - Online Fairness-Aware Learning Under Class Imbalance | 2020 | Springer International Publishing | Vasileios Iosifidis, Eirini Ntoutsi |
| 74 | FairBatch: Batch Selection for Model Fairness | 2021 | Not Specified | Yuji Roh, Kangwook Lee, Steven Euijong Whang, Changho Suh |
| 75 | Fairness via Representation Neutralization | 2021 | Curran Associates, Inc. | Mengnan Du, Subhabrata Mukherjee, Guanchu Wang, Ruixiang Tang, Ahmed Hassan Awadallah, Xia Hu |
| 76 | Gradient Driven Rewards to Guarantee Fairness in Collaborative Machine Learning | 2021 | Curran Associates, Inc. | Xinyi Xu, Lingjuan Lyu, Xingjun Ma, Chenglin Miao, Chuan Sheng Foo, Bryan Kian Hsiang Low |
| 77 | Exploiting Transferable Knowledge for Fairness-aware Image Classification | 2020 | Springer International Publishing | Sunhee Hwang, Sungho Park, Pilhyeon Lee, Seogkyu Jeon, Dohyung Kim, Hyeran Byun; |
| 78 | Joint Transfer of Model Knowledge and Fairness Over Domains Using Wasserstein Distance | 2020 | IEEE | Taeho Yoon, Jaewook Lee, Woojin Lee |

Table 11 (continued)

| ID | Title | Year | Publisher | Author |
|---|---|---|---|---|
| 79 | Individual Fairness for Graph Neural Networks: A Ranking based Approach | 2021 | Association for Computing Machinery | Yushun Dong, Jian Kang, Hanghang Tong, Jundong Li |
| 80 | Enhancing Long Term Fairness in Recommendations with Variational Autoencoders | 2019 | Association for Computing Machinery | Rodrigo Borges, Kostas Stefanidis |
| 81 | Neural-Symbolic Integration for Fairness in AI | 2021 | Not Specified | Benedikt Wagner, Artur d'Avila Garcez |
| 82 | Online Set Selection with Fairness and Diversity Constraints | 2018 | OpenProceedings.org | Julia Stoyanovich, Ke Yang, HV Jagadish |
| 83 | Imparting Fairness to Pre-Trained Biased Representations | 2020 | IEEE | Bashir Sadeghi, Vishnu Naresh Boddeti |
| 84 | Subgroup Generalization and Fairness of Graph Neural Networks | 2021 | Curran Associates, Inc. | Jiaqi Ma, Junwei Deng, Qiaozhu Mei |
| 85 | Achieving Differential Privacy and Fairness in Logistic Regression | 2019 | Association for Computing Machinery | Depeng Xu, Shuhan Yuan, Xintao Wu |

**2.2.6 Scopes and objectives of the questions.** In this section, we will discuss about our purpose of selecting these questions from the fair machine learning-related question space, the scope of obtaining answers to the questions from the selected articles, and what the outputs of these questions can contribute to the user. Table 12 shows the answers of the selected questions about each article in the selected articles. Parts are written as "Not Specified" which means that no answer could be found to the relevant question in the article.

In the first question, we examined in which areas the problems created by the bias in protected features, which is the reason for the emergence of fair artificial intelligence models, are focused. In this way, researchers and institutions can create an awareness of the areas and tasks that fair AI should focus on but have not been addressed so far. Furthermore, researchers can learn from the output of the relevant question in which areas they should develop a fair artificial intelligence system for the problems. In this question, we aimed to raise awareness by removing the focal points of future researchers and focusing on certain issues for researchers who do not have knowledge about which areas to focus on problems based on this knowledge. In the articles, we took the areas and problems mentioned in the definition of artificial intelligence as answers.

In the second question, we investigated which fairness metrics were used by the models proposed in the selected articles. With the help of this question, it can be determined whether group fairness or individual fairness is used more. In addition, it can be examined which rate is more important for fair artificial intelligence from the fairness metrics proposed by TPR, FPR, TNR, and FPR calculations. Based on this question, we collected the metrics that were previously defined, suggested by the user, or suggested as fairness metrics, which are not directly fairness metrics, as answers.

In the third question, we investigated which machine learning models are used for fair AI. Since the articles belong to all three equitable artificial intelligence methodologies (pre-, in-, and post-processing), we preferred to show the general trend instead of presenting which one is most used in each methodology in the selected articles. With the answer to this question, researchers can access which machine learning models they prefer for the fair AI model or the proposed bias reduction system, or which models are not currently used. In the articles, the models used for the experiment and the models chosen as a fairer version of an existing machine learning model are addressed as answers.

Finally, the last question addresses the challenges faced by fair AI. We hope that this question will guide the models that can be proposed to further develop fair artificial intelligence for future research. Related to this question in the articles, we have addressed the points that are stated as difficulties in the study, such as difficulties

related to artificial intelligence ontology, difficulties related to the definition of fairness, and problem-based difficulties, which are presented to the readers as future research aims.

Table 12

*Answers of the Questions in Selected Publications*

| ID | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|
| 1 | Classification | Equalized Odds | Auto-encoder block + classification block | Not Specified |
| 2 | Multi-business | Hirschfeld-Gebelein-Renyl (HGR) | Simple neural network | Fairness decreases the accuracy of the model.<br>Can cause negative impact on some individuals. |
| 3 | Decision-making systems | Statistical parity | Random forest | Not Specified |
| 4 | Network representation | Structural heterogeneity<br>Mean latent similarity discrepancy (MLSD) | K-means | Missing ground truth can cause evaluating problems on accuracy and fairness |
| 5 | Face Recognition | Demographic Parity | VGGNet-16 (Configuration D) | Not Specified |
| 6 | Retinal Diagnostics and Rare Ophthalmic Diseases | Equalized odds<br>Equal accuracy | ResNet50 | Different approaches can be needed for different retinal diseases.<br>There may be other fairness requirements that are not well captured by those metrics but that merit use. |
| 7 | Facial Image-to-Image translation | Fréchet Protected Attribute Distance | Generative adversarial networks | Not Specified |
| 8 | Decision-making systems (Multiple) | Equalized Odds | Adaptive rejection learning | Machine learning approaches to fairness is formulating an operational definition |

Table 12 (continued)

| ID | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|
| 9 | Decision-making system | Demographic parity | Generative adversarial networks | Not Specified |
| 10 | Data mining | Statistical parity | Generative adversarial networks | Achieving equalized odds or equal opportunity on the besides statistical parity will be investigated |
| 11 | Prediction systems | Flexible fairness parity | Simple learning algorithm using p | It may be further interest within the study of model interpretability |
| 12 | Classification and intervention task | Accuracy gap | CFMLP, CF4MLP CVAE-A FCVAE | Classification suggests the more equitable deployment of machine learning when only biased data are available but also raises significant technical challenges. |
| 13 | Decision-making systems | Equal Opportunity Equalized Odds | Random Forest | Considering richer utility functions relevant to real-world decision systems |
| 14 | Ranking applications | Statistical Parity Disparate Exposure Disparate Impact | Policy-gradient algorithm | Not Specified |
| 15 | Decision-making systems Risk assessment Medical predictions | Demographic Parity Error Parity False Discovery rate Parity False Negative rate Parity | Amazon Mechanical Turk | One barrier to obtaining meaningful answers from participants is engagement. As with any experiment, they cannot completely rule out the potential impact of framing. |

Table 12 (continued)

| ID | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|
| 16 | Decision-making systems | Contrastive fairness | Neural network | However, most causal questions of fairness in real life are contrastive and many contrastive questions are asked at individual level. The assumptions of linearity and independence that may hold true at population level might not be valid per individual. |
| 17 | Decision-making systems | Burden | Genetic algorithm<br>Inception-v3<br>ResNet-50<br>MobileNet<br>ImageNet | Not Specified |
| 18 | Robotics | Procedural fairness | Strict-rate-constrained UCB Algorithm<br>Stochastic-rate-constrained UCB Algorithm<br>Amazon Mechanical Turk | Generalizability of findings to human-robot teamwork with physically embodied robots |
| 19 | Automation | Expectation Fairness (EF)<br>Strict Fairness | Neural architecture search<br>NSGA-II<br>ImageNet | Not Specified |
| 20 | Decision-making systems<br>Stream classification | Statistical parity | Hoefdding Trees<br>Accuracy Updated Ensembles<br>Naive Bayes<br>k-Nearest Neighbors | The effect of “data correction for discrimination” on a variety of classifiers is different, and therefore, how to “best correct” for specific classifiers is an open question |

Table 12 (continued)

| ID | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|
| 21 | Decision-making systems<br>Data stream classification | Statistical parity | Hoeffding Tree | Not Specified |
| 22 | Decision-making systems | Statistical parity<br>Disparate impact<br>Normalised prejudice index | Decision Trees<br>Random Forests<br>(Pre-processing) | Removing attribute does not always lead to models that make fairer predictions. Researchers believe this behaviour is caused by indirect prejudice, due to the presence of other attributes highly associated with the sensitive attribute. |
| 23 | Classification | Individual fairness | Logistic regression | Principled standard for ensuring subpopulation fairness will be investigated |
| 24 | Medical decision-making systems | True positive rate<br>Equal opportunity | Gradient boosting trees | Allocating resources using 'biased' models may worsen health disparities |
| 25 | Computer vision | Not Specified | Class incremental learning (CNN) | Not Specified |
| 26 | Computer vision | Test accuracy | ResNet-50 | What happens if the imbalanced domain distribution is not known at training? |
| 27 | Decision-making system | Demographic parity<br>Equal opportunity<br>Equalized odds | MLP (Semi-supervised) | Not Specified |
| 28 | Decision-making system | Equal opportunity | Logistic Regression | Not Specified |

Table 12 (continued)

| ID | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|
| 29 | Decision-making systems<br>Downstream applications | Equalized odds<br>Equal opportunity<br>Statistical parity difference | Asynchronous advantage actor-critic structure learning (Reinforcement learning) | Not Specified |
| 30 | Decision-making systems<br>Medical diagnostics | Not Specified | Neural network<br>Logistic Regression<br>ResNet-18 | It requires a target task to learn the disentanglement which could be avoided by learning the reconstruction as an auxiliary task |
| 31 | Multi-object tracking | Not Specified | CenterNet | Not Specified |
| 32 | Recommender systems | Item index | Collobrative filtering | It is more difficult to increase recommendation fairness when demographic data is missing |
| 33 | Decision-making system | Demographic parity<br>Equal opportunity | Bayes-optimal classifier | Extensions beyond linear metrics, to consider more general fractional and convex metrics |
| 34 | Decision-making system | Demographic parity<br>Equalized odds | Meta algorithm (Classifier) | Perform a similar fairness vs robustness analysis for other notions of fairness. |
| 35 | Facial expression recognition | Equal opportunity | ResNet-18 | Requires explicit attribute information apriori (e.g., the age, gender and race of the subject whose expression is being classified) which may not be easy to acquire in real-world applications. |

Table 12 (continued)

| ID | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|
| 36 | Federated learning | Coefficient (Collaborative fairness) | Federated Learning | How to quantify fairness in more complex settings, and apply their framework to various domains, such as financial, biomedical, speech, NLP, etc.<br>Furthermore, researchers would like to systematically integrate robustness with fairness. |
| 37 | Face recognition | Mean distance | DEX-VGG | Addressing the fairness of age prediction systems is challenging. |
| 38 | Decision-making systems | Discrimination score | Probabilistic circuits | Not Specified |
| 39 | Decision-making systems | Demographic parity | Logistic regression (Semi Supervised Learning)<br>Support vector machine (Semi Supervised Learning)<br>Ensemble learning | How to achieve fair semi-supervised learning where labeled and unlabeled data have different data distributions |
| 40 | Natural language processing | Empirical privacy | Differentially Private Neural Representation (BERT) | Extensions beyond linear metrics, to consider more general fractional and convex metrics |

Table 12 (continued)

| ID | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|
| 41 | Decision-making system | Equal opportunity | Nondominated Sorting Genetic Algorithm II (Active Learning) | Decision-maker cannot be made available on-demand nor altered, the true labels of historical data are unavailable, only a subset can be acquired, and the augmented inferences must be easily understandable by humans. |
| 42 | Clinical natural language processing | Equalized odds | Channel-wise bidirectional LSTM<br>CNN<br>Logistic regression (Ensemble Learnig) | Demographic parity may be unfair if the underlying group-specific base rates are unequal<br>They encourage further discussion into concretely defining fair, real-world NLP tasks and developing novel algorithms. |
| 43 | Anomaly detection | Statistical parity<br>Group fidelity | Deep-autoencoder | Not Specified |
| 44 | Skin segmentation | Equalized odds | Generative adversarial networks | Adversarial independence may still produce biased results with regard to a protected factor that was not tested against or has yet to be considered protected. |
| 45 | Classification | Equal opportunity | LSTM | Maximizing the distortion factor is a challenging optimization problem.<br>Throughout their analysis they assume that the datasets D1, D2 are sampled from the same distribution therefore extend their theory to the more realistic case of when the dataset distributions are |

Table 12 (continued)

| ID | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|
| | | | | different would be interesting. |
| 46 | Social networks topologies | Demographic parity<br>Equalized odds | Graph neural network (Graph Representation Learning) | A possible limitation of FairDrop is its extension to the case where multiple sensitive are present at once.<br>Randomness is a fundamental concept for FairDrop. However, not every connection has the same importance inside a graph. |
| 47 | Feature selection | Proportion as a compliment since Balance | Fairness-aware Unsupervised Feature Selection | Despite its fundamental importance, the fairness of unsupervised feature selection has largely remained nascent. |
| 48 | Multi-task learning | Equal opportunity<br>Equalized odds<br>Average relative fairness gap<br>Average relative error | Multi-taskaware fairness treatment | Fairness in multi-task learning poses new challenges and the need of characterizing a multi-dimensional Pareto frontier. In a traditional multi-task learning setting where fairness is not taken into consideration, people focus on optimizing the Pareto frontier of multiple accuracies across tasks. Instead, their work aims at demystifying the multi-dimensional trade-off and improving fairness on top of accuracy objectives for multi-task learning problems. |
| 49 | Face recognition | False positive rate (FPR)<br>False negative rate (FNR) | ResNet34<br>ResNet50<br>ResNet100 | Designing a better weight function for inconsistency penalty, and investigating the effects of noise samples that might be mistakenly optimized as false positive cases. |

Table 12 (continued)

| ID | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|
| 50 | Medical image segmentation | Standard deviation<br>Skewed error ratio<br>Average dice similarity coefficient | nnU-Net<br>Meta-learning (DenseNet) | Current clinical practice the outputs of such models are typically manually modified by a clinician. However, this modification process is time-consuming and prone to error, so does not eliminate the bias completely |
| 51 | Recommender systems | Weighted proportional fairness | Actor-critic framework (Reinforcement learning) | Not Specified |
| 52 | Online learning | Decision Boundary Covariance<br>Demographic Parity<br>Equalized Odds<br>Discrimination | Follow-the-fair-meta-leader (Online meta learning) | It remains interesting if one can prove that fairness constraints are satisfied at each round without approximated projections onto the relaxed domain, and if one can explore learning when environment is changing over time. |
| 53 | Decision-making system (Text Style Obfuscation) | True Positive Rate | VAE | One potential issue they see is the chance that systems like this might obfuscate text by converging towards the majority and erasing styles of marginalized communities. |
| 54 | Natural language processing | Equalized odds | IW<br>INLP<br>LDAM<br>LDAMcw (vanilla class-imbalanced learning) | More complex tasks and multiple private attributes |

Table 12 (continued)

| ID | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|
| 55 | High-stake applications | Equal opportunity<br>Demographic parity | Logistic Regression | Automatic methods to identify causal and spurious features and extend the regularization approach to complex deep learning frameworks. |
| 56 | Clustering algorithms | Balance fairness | k-median (Adversarial attack) | Not Specified |
| 57 | Decision-making systems | Disparate impact<br>Disparate mistreatment | One-class support vector machine | How a pre-trained network adapts to the new dual objective |
| 58 | Decision-making systems | Path-specific counterfactual fairness | Latent inference-projection | Not Specified |
| 59 | Decision-making systems | Demographic parity<br>Equalized odds<br>Weighted proportional fairness | Logistic regression | Not Specified |
| 60 | Decision-making systems | Equal opportunity<br>Equalized odds<br>Equalized accuracy | Variational Auto-Encoder | Not Specified |
| 61 | IoT | Accuracy | Cluster-based Federated Averaging (Hierarchical Clustering with K decision) | Changing the fixed weight to a dynamic one in a low-cost energy, and extending the framework to accept new devices/clients dynamically. |

Table 12 (continued)

| ID | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|
| 62 | Decision-making systems | Disparate treatment elimination<br>Disparate mistreatment elimination | Binary probabilistic classifier | Statistical models often arrive at a minimum condition that guarantees correct but not necessarily full treatment of training bias.<br>Inability to justify disparate mistreatment elimination. |
| 63 | Recommender systems | Absolute difference between mean ratings of different groups<br>Kolmogorov-Smirnov statistic | Gradient Descent<br>Newton's Method<br>(Fairness-Aware Tensor-based Recommendation) | Generalizing their framework to consider alternative notions of fairness beyond statistical parity. By extending their framework in this direction<br>How to incorporate real-valued features into the framework for recommenders with explicit ratings, and in running user studies on the perceived change of fairness |
| 64 | Decision-making systems | Demographic parity<br>Equalized odds<br>Equal opportunity | K-shot Fairness (Model Agnostic Meta Learning) | Receiving a non-unfair score in fairness warnings does not guarantee that the model will behave fairly in the new domain |
| 65 | Datasets | Causal fairness | Generative adversarial network | Not Specified |
| 66 | Decision-making systems | Balanced error rate | Generative adversarial network | A more powerful classifier exists that could infer the sensitive attribute with higher accuracy. Note that this is an inherent limitation of all the preprocessing techniques and not only their approach.<br>Assess the relationship between the different fairness notions, namely the |

Table 12 (continued)

| ID | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|
| | | | | impossibility of inference and the individual and group fairness. |
| 67 | Robotics | Distributive fairness<br>Linear Temporal Logic<br>Demographic parity | Pareto estimation method | Ethics of risk imposition, and the social and ethical implications of apparently innocent behaviour of autonomous systems are important topics to further explore. Such reflections are necessary for us to better align our robots with our values, and to better anticipate the impact autonomous systems have in society and our lives. |
| 68 | Recommender systems | Individual fairness | Decision tree ensembles | Fairness verification and learning method by considering alternative fairness definitions, such as group or statistical fairness, or stronger notions such as causal or dependency fairness. Designing quantitative verification methods for both stability and fairness. |
| 69 | Decision-making systems | Probabilities | Optimal offline algorithm | As in many real-world settings the online decisions go beyond the single selection model studied here, there is ample opportunity for extending this line of work to combinatorial settings. |
| 70 | Natural language processing | Equalized odds<br>Equality opportunity | GloVe-based CNN<br>BERT | Effects of robustness and fairness in attributes other than gender and sexual orientation, extending our study to other word |

Table 12 (continued)

| ID | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|
| | | True Positive Equality Difference<br>False Positive Equality Difference | | substitution based robustness methods, and exploring more sophisticated methods to combine robustness and bias mitigation methods during training. |
| 71 | Decision-making systems | Information gain | Hoeffding Tree | A different avenue is to extend these results in conjunction with their previous work to situations where the class label is not available for fair clustering |
| 72 | Automated systems | Absolute odds difference | Classifier | Not Specified |
| 73 | Decision-making systems | Cumulative statistical parity<br>Cumulative equal opportunity | Adaptive Hoeffding Trees (Class imbalance-aware boosting) | Embed the decision boundary adjustment directly into the training phase by altering the weighted training distribution. |
| 74 | Sensitive applications (healthcare and finance) | Equalized odds<br>Equal opportunity<br>Demographic parity | Logistic regression (Batch selection algorithm) | Not Specified |
| 75 | High-stake applications | Demographic Parity<br>Equalized Odds | ResNet-18<br>MLP (Representation Neutralization for Fairness) | Learning debiased representations is a technically challenging problem. |
| 76 | Federated learning | Cosine Gradient Shapley Value | CNN (Federated learning) | Can be achieved both optimally or is there some inescapable trade-off between fairness and performance? |

Table 12 (continued)

| ID | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|
| 77 | Computer vision | Equality opportunity<br>Group-wise Fair loss | Protected Attribute Classifier<br>Target Attribute Classifier (Transfer Learning) | Transfer between different domains is not investigated in this work. Adopting domain adaptation techniques would be interesting for future work. |
| 78 | Decision-making systems | Strong Pairwise Demographic Disparity<br>Strong Pairwise Disparity of Opportunity<br>Demographic parity<br>Equal Opportunity | Not Specified | Could be extended to more intricate tasks pertaining to domain adaptation and fair machine learning. Could be generalized for different notions of fairness, including individual fairness. |
| 79 | Decision-making systems | NDCG@k<br>ERR@k<br>(Individual fairness) | Graph neural network | Besides, REDRESS is not only restricted on GNNs but can be extended onto other graph mining models and tasks. |
| 80 | Recommender systems | Distance between attention and relevance distributions | Variational Auto-Encoder | Specifically of having a list to be presented as a recommendation where the items in the first positions have the same or very similar relevance |
| 81 | Knowledge representation | Demographic parity<br>Disparate impact | Logic Tensor Networks | Suitability of different XAI methods. Varying notions of fairness as the proposed method itself is adaptable and there are lively discussions about the appropriateness of varying definitions. |

Table 12 (continued)

| ID | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|
| 82 | Decision-making systems | Normalized difference<br>Elift ratio | Online Algorithm | Not Specified |
| 83 | Image representation | Not Specified | Adversarial Representation Learning | Not Specified |
| 84 | Graph-level tasks | False Positive Rate (FPR)) | Graph neural network | In practice, the fairness issues of learning on graphs are much more complicated due to the asymmetric nature of the graph-structured data. Analyze the fairness of GNNs on real-world applications and develop principled methods to mitigate the unfairness. |
| 85 | Decision-making systems | Risk difference | Logistic regression | Allocation strategies of privacy budget, e.g., adding different amount of noise to coefficients containing different attributes. |

**2.2.7 Results.** In this section, the results of our systematic literature review will be presented. Selected papers by the year are illustrated in Figure 7. The publications published in 2021 constitute the majority of the publications we selected.

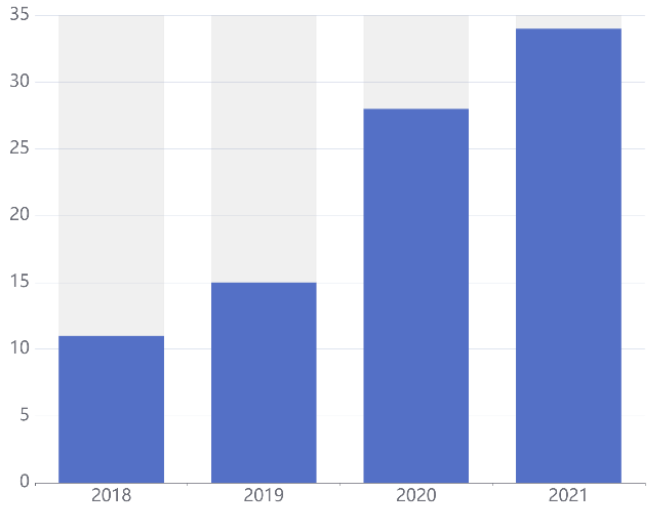


*Figure 7.* Distribution of publications by year

Furthermore, selected papers by the publishers illustrated in Figure 8 Association for Computing Machinery published 19 papers of 85 papers. Springer International Publishing and Curran Associates Inc. published 13 and 11 respectively. 7 of the publications have been published by IEEE and both AAAI Press and MDPI have published 6 publications. 23 of the publications as categorized as "Others/Not published yet" which means publications are not published yet or have been published by publishers such as IBM, BMVA Press, and Elsevier that have published fewer articles about this term.

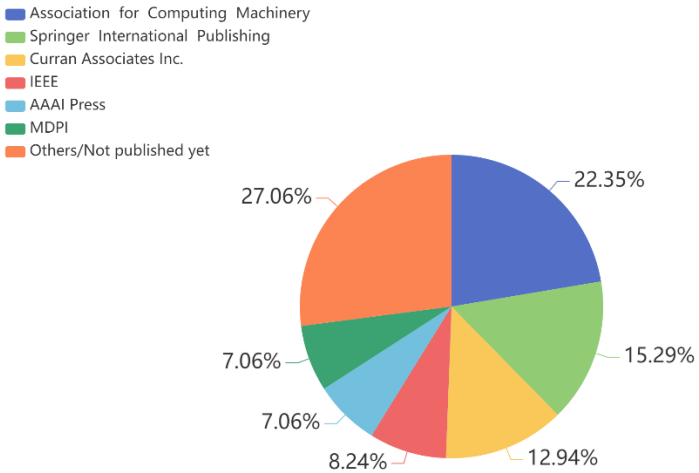


Figure 8. Distribution of publishers

The four research questions presented in Table 5 are addressed one by one in the following subsections:

#### *2.2.7.1 RQ1: What are the areas where fair artificial intelligence has been used?*

- Decision making systems: Decision-making systems have been started to use in complex domains such as loan approvals, credit card approvals, and criminal justice has raised questions about the role of machine learning in high-stakes decision-making(“Machine Bias — ProPublica,” n.d.) systems (Madras, Pitassi, & Zemel, 2018; Noroozi et al., 2019). With this trend, fairness became a metric that should be taken into account for decision-making machine learning models. COMPAS algorithm is an example of racial bias in the criminal justice decision-making system. According to ProPublica, Black defendants were presumed to be at greater risk of recidivism than they were.
- Computer vision: Computer vision cases notably in face recognition have some insulting or biased-discriminated results based on protected attributes. The most known example of the bias in computer vision is labeling black people as "gorillas" by Google Photos (“Google Apologises for Photos App’s Racist Blunder - BBC News,” n.d.). This insulting example shows that models can be biased from skin color. Another problem in computer vision is causing unprotection of privacy. Machine models can simultaneously extract useful information from data and can reveal protected information of individuals (Franco, Oneto, Navarin, & Anguita, 2021).
- Medicine: Deep learning (DL) models have shown success in many medical image segmentation tasks (Puyol-Antón et al., 2021). Mechanisms leading to cardiovascular disease, vary according to demographic characteristics such as race and gender. Therefore, fair machine learning models play a crucial role in disease detection. And also, medical data can cause bias based on the distribution of samples. An interesting example of this problem is AREDS (Age-Related Eye Disease Study) (Kassoff et al., 2001). In this study in which few participants older than 85 years are present, the automated diagnoses might be biased against correct diagnoses for individuals older than 85.

- Natural language processing: NLP has sub-branches that are affected by bias as indicated by studies. Text representation (Lyu, He, & Li, 2020), is one of these branches and an individual's input text can contain some protected information such as gender, age, and other important attributes and NLP methodologies can reveal them (Preot'iuc-Pietro, Lampos, & Aletras, 2015). Another interesting NLP task that can be affected by bias is class imbalance. Class imbalance can be defined as a high correlation of labels with one or more protected features. For example, more men are employed as engineers than women and that can cause a bias towards individuals.
- Recommendation systems: Bias in traditional recommendation systems, has attracted increasing attention in studies (Bobadilla et al., 2021a; Borges & Stefanidis, 2019; W. Liu et al., 2020). In most the recommendation systems, items are listed in order from which one item is selected. Therefore, users tend to pay more attention to the items in first positions, instead of higher rankings. Items which have equal or close probabilities should have been shown to the users in proper order, in other words, items should be sorted by their scores. Due to bias, high-scored items can be less preferred owing to their order in the item list

***2.2.7.2 RQ2: Which fairness definitions have been taken into account for fair artificial intelligence?*** While there are more than 70 metrics have been proposed as fairness metric, researchers have been focused on more specific metrics than others. As shown in Figure 9, demographic parity is the most used fairness metric in the primary papers with 23.53% of the papers. Equalized odds and equal opportunity follow the demographic parity with 16.47% and 15.3% respectively. Disparate impact is used less than other metrics with 5.88%. While 33 primary studies are categorized as "Other/Not specified", most of these papers include uncommon fairness metrics such as Hirschfeld-Gebelein-Renyl (Grari et al., 2021), structural heterogeneity (X. Du, Pei, Duivesteijn, & Pechenizkiy, 2020) or metrics can be used to measure fairness in an indirect way such as accuracy (F. Zhao, Huang, Maradapu Vera Venkata Sai, & Wu, 2020), or they have not specified metrics in their papers**.**

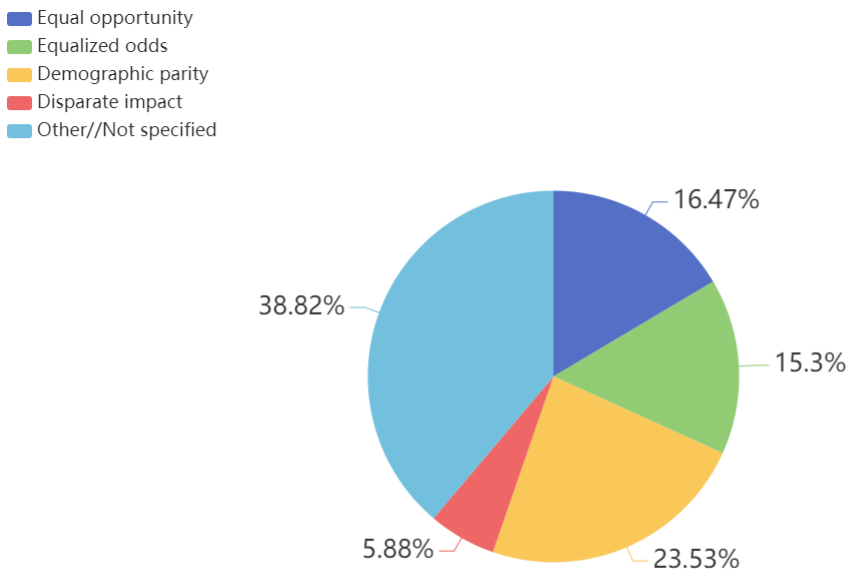


*Figure 9.* Distribution of fairness metrics

By looking at the percentages of use of the metrics, we can see that group fairness is preferred more than individual fairness in studies. A few reasons for this situation are as follows:

- Individual fairness focuses on similar individuals getting similar outputs, and it can be difficult to find out which features should be used to infer from this similarity function.
- Noise and mislabeling can greatly distort outcomes in individual fairness.
- It can increase group-based unfairness even more because individuals in disadvantaged groups who need to get positive results may not get these outputs with the use of the characteristics that need to be protected in individual fairness.

For these reasons, we can say that group fairness and individual fairness are inversely proportional, and if we look at the current problems of prejudice and the examples we have presented in this study, we can say that prejudice is based on group-based characteristics rather than individuals. This inference includes an inference about whether individual fairness should be taken into account in the research rather than making a judgment about whether or not individual fairness is unimportant.

Demographic parity, equalized odds, equal opportunity, and disparate impact, which are binary-classification-based group fairness metrics, have been used more than other group fairness metrics, due to a large number of applications or examinations of decision-making systems based on the classification in the field of fair artificial intelligence. One of the most important reasons for this is that COMPAS, German Credit, Law School, and Adult Census datasets, which are the most used in fair artificial intelligence, are binary classification problems or can be converted to them. Among these first four metrics, one of the reasons why demographic parity, the first presented as a fairness metric, is used the most is that it is simpler than other metrics and its conditions are easier to implement. According to demographic parity, the positive output rates of the groups should be equal to each other and this is a method that can be used to eliminate the unfairness of the unprivileged group.

Equalized odds, which come after demographic parity, can be defined as the closest and most restricted fairness metric to today's understanding of fairness. Therefore, equalized odds are one of the most appropriate metrics for researchers who want to create a fair artificial intelligence model close to everyday life among the fairness metrics. According to equalized odds, protected and unprotected groups have equal true positive rates and equal false positive rates. Equal opportunity can be defined as a more relaxed version of equalized odds. Equal opportunity states that each group should get the same true positive rate.

***2.2.7.3 RQ3: Which methodologies have been proposed as a fair artificial intelligence model?*** Researchers have proposed fair machine learning techniques as a part of fair artificial intelligence term. By specifying the derivatives of the models as the main model, we prevented the calculation error that may occur from sub-branches. For example, we have categorized NSGA-II as a genetic algorithm in this case. Figure 10 shows the proposed machine learning models with this conversion. Logistic regression is the most used machine learning model with 11.76% and it is followed by decision-tree-based models (Hoeffding tree, random forest, and so on) pointed at 10.59%. 8,23% of published papers has used ResNet-based architectures (ResNet50, ResNet100, and so on) and both autoencoder-based models and generative adversarial networks have been used by 7,06%. Likewise, CNN and MLP have been used by 4.71% and genetic algorithms (NSGA-II for example) had less usage in 3,53%

compared with these machine learning models. While majority of the methods are categorized as "Other\Not Specified", this category includes less used methodologies such as asynchronous advantage actor-critic structure learning (A3SL) (Yue Zhang & Ramesh, 2020), probabilistic circuits (Choi, Dang, & van den Broeck, 2021a) or not defined in detail. There are several reasons why logistic regression is most commonly used. First of all, it is a suitable machine learning model for adding the fairness metric used for the in-processing phase to the loss function. Secondly, logistic regression can solve the aforementioned problems with few parameters and can directly observe the effect of fairness metrics on metrics such as the accuracy of the model. Finally, logistic regression can be used in learning methodologies that are most used by decision support systems namely supervised and semi-supervised learning. For example, (T. Zhang et al., 2022) used logistic regression in semi-supervised settings, (Zhao Wang, Shu, & Culotta, 2021) used in supervised settings.

Two models that can be considered very important here, GAN and decision tree, have taken their place in the models frequently used in fair artificial intelligence models. The importance of these two models for us is the fair data set generation provided by GAN and the high explainability provided by the decision tree models. Data quality and the fact that the data are free from biases are of great importance for future models. The studies carried out by the European Union regarding this importance are also instructive (– European Union Agency for Fundamental Rights, n.d.; “4 Possible Ways to Avoid Big Data Bias | European Union Agency for Fundamental Rights,” n.d.; “Addressing Algorithmic Discrimination in the European Union - A Path For Europe (PfEU),” n.d.; “Open Data and Data Bias | Data.Europa.Eu,” n.d.). GAN is one of the most important models that can be used to create fair data. It is one of the options that can guide the future studies that GAN's ability to create a fair data set instead of only the classification problem in fair artificial intelligence is emphasized future studies. The decision tree, on the other hand, is one of the highly explainable machine learning models, including models that are frequently used in ensemble learning. Demonstrating fairness on the basis of explainability rather than just satisfying fairness metrics will be important for further studies to expand the definition and dynamics of fairness.

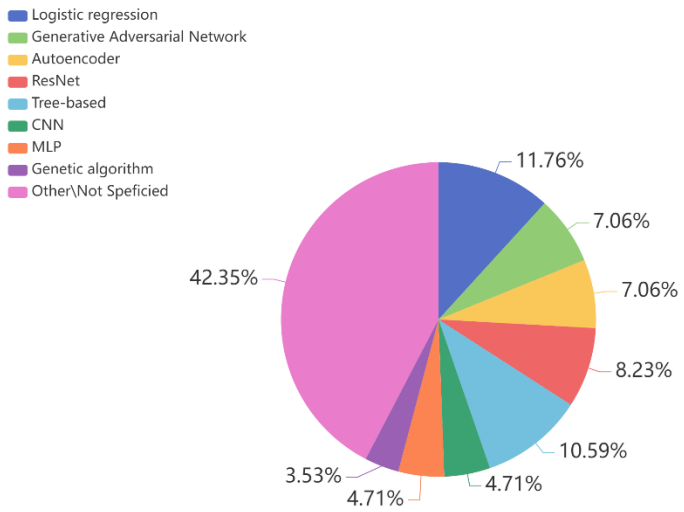


*Figure 10.* Distribution of fair machine learning models

***2.2.7.4 RQ4: What are the drawbacks, challenges, and possible solutions in fair artificial intelligence?*** We categorized the challenges into five categories. Table 13 presents these categories. These five categories are described as follows:

- Extendability: There are more than 70 fairness metrics which has been proposed and it is impossible to satisfy all fairness metrics for fair artificial intelligence models. Because there are trade-offs between some fairness metrics in mathematical notions. For example, there is a conflict between statistical and individual fairness, and in the COMPAS criminal recidivism risk scoring algorithm, it was proven that it is impossible to simultaneously achieve two group-fairness measures; namely, false positive equality and equal calibration (Chouldechova, 2017). To handle this issue, a different perspective has been proposed by (Srivastava, Heidari, & Krause, 2019) by running several experiments to investigate the most compatible notion of fairness with each participant's choices.
- Interaction: (Madras et al., 2018) points out that in many machine learning applications, there is more than one decision-maker involved, both algorithm and human however, in algorithmic design, their interactions with each other are often unaddressed. To address this problem, they have proposed a two-stage framework that contains an automated model and an external decision-maker. They also proposed learning to defer, which is a generalization of the rejection

learning by considering the effect of other agents in the decision-making process.

- Real-world differences: Fair artificial intelligence models focus on mitigating bias in protected features in most cases. However, while fair artificial intelligence algorithms focus on counterfactual questions, such as “what if?” or “why?”, real-world problems consist of contrasting consequences of subjective questions such as "why this but not that?”. (Chakraborti, Patra, & Noble, 2020) handles this question and proposed new algorithmic fairness named” contrastive fairness” which is comparative fairness.
- Machine learning process: (Iosifidis, Fetahu, & Ntoutsi, 2019) focused on the problem of existing fair artificial intelligence models that is focusing solely on the pre-, in-, or post-processing steps instead of covering all steps in their models. In their perspective, the fairness problem cannot be solved only in a single step of the machine learning process thus holistic approach is required to solve this issue. In the light of this perspective, they have proposed FAE (Fairness-Aware Ensemble) framework that combines fairness-related interventions at both pre- and post-processing steps in the machine learning process.
- Trade-off between fairness and accuracy: Existing proposed fair machine learning models have increased fairness while sacrificing accuracy at the same time. The trade-off between fairness and accuracy comes up as a challenge for researchers while developing fair machine learning models. To solve this problem, (T. Zhang et al., 2022) proposed a fair semi-supervised learning framework benefiting from that increasing the size of the training set may lead to a better trade-off.
- Domain adaptation: One of the biggest problems in fair artificial intelligence is that the proposed fair artificial intelligence is limited to the existing data set or problem. Therefore, domain adaptation techniques are very important for the implementation of fair artificial intelligence models in real-world systems (Hwang et al., 2021).

Table 13

*Challenges and Solutions for Fair Artificial Intelligence*

| Category | Challenges (C1 to C6) | Solutions (S1 to S6) | Reference |
|---|---|---|---|
| Extendability | It is impossible to satisfy all fairness metrics for fair artificial intelligence models | Running several experiments to investigate the most compatible notion of fairness with each participant's choices | (Srivastava et al., 2019) |
| Interaction | In algorithmic design, multiple decision-maker interactions with each other are often unaddressed | Proposing a two-stage framework that contains an automated model and an external decision-maker and learning to defer by considering the effect of other agents | (Madras et al., 2018) |
| Real-world | Real-world problems consist of contrasting consequence of subjective questions instead of counterfactual questions | Proposing comparative fairness measurement | (Chakraborti et al., 2020) |

Table 13 (continued)

| Category | Challenges (C1 to C6) | Solutions (S1 to S6) | Reference |
|---|---|---|---|
| ML process | Existing fair machine learnings model that is focusing solely on the pre-, in-, or post-processing steps instead of covering all steps | Proposing framework that combines fairness-related interventions at both pre- and post-processing steps in the machine learning process | (Iosifidis, Fetahu, et al., 2019) |
| Trade-off | Fair machine learning models have increased fairness while sacrificing accuracy. | Increasing the size of the training set may lead to a better trade-off between fairness and accuracy | (T. Zhang et al., 2022) |
| Domain adaptation | Current fair machine learning models are task or data-specific. | Domain adaptation methods should be included as part of fair artificial intelligence methods. | (Hwang et al., 2021) |

**2.2.8 Discussion of systematic literature review.** In this work, we reviewed the literature on fair artificial intelligence to understand the state-of-the-art and current practices. For this purpose, four research questions were identified and investigated.

RQ1 aimed at discovering the areas for involving fairness in artificial intelligence. Bias in these areas namely decision-making systems, computer vision, medicine, natural language processing, and recommendation systems were the top motivations respectively. This finding shows us the bias in high-stake applications directs researchers to involve fairness in artificial intelligence.

RQ2 focused on which fairness definitions have been applied in artificial intelligence. Demographic parity, equal opportunity, equalized odds, and disparate impact are the most used fairness metrics among all metrics. As shown group fairness is more taken into account than individual fairness in fair artificial intelligence models.

RQ3 focused on machine learning models which are used as a part of fair artificial intelligence. Logistic regression, decision-tree-based models, ResNet, autoencoders, GANs, CNNs, and MLPs are the most used machine learning methodologies in fair artificial intelligence. Researchers tend to use classification models and neural networks in fair artificial intelligence methods.

RQ4 identified the key challenges and possible solutions faced by prior researchers. The collected challenges were mainly the challenges of building machine learning methodologies that are unaddressed real-world dynamics. Challenges were reported based on the explicit statements in the articles. There can be more challenges, however, if they have not been mentioned in these papers, we could not identify and include them here. We can see new approaches and challenges when fair artificial intelligence models are found in wider usage in the real world. There are several threats to the validity of this SLR. Concerning the time frame, the primary papers selection process was finalized in January 2022. Microsoft Academic is unavailable after Dec. 31, 2021 therefore reaching articles using Microsoft Academic will be impossible after this date. We have collected publications from there before it goes inactive. We have used article search engines and we have eliminated non-peer-reviewed articles. Also, the search for the primary papers was strictly focused on papers in English, as such, there could be a chance of missing some papers that were written in other languages that could add value to the research questions in this paper.

Finally, for the development of fair AI and for real-world systems to be fairer, issues such as unbiased data use, explainability, and domain adaptation should be more associated with fair AI by researchers. Although the majority of these studies are based on group-based fairness, in which individuals are grouped according to the characteristics that need to be protected, the elimination of individual-based unfairness should be examined at least as much as group fairness in future studies.

# Chapter 3: Methodology

In this section, we will address the existing problems and the methodology we propose on these problems.

## 3.1 Research Design

In this section, we will present our new method, experiment architecture, and information for used methodologies.

**3.1.1 Hybrid fairness.** We used in-processing as the fair artificial intelligence methodology and added the fairness as a parameter to the loss function in in-processing. This fairness function is also defined as a combination of group fairness and individual fairness. Our total loss function can be defined as in Equation 11:

$$\boldsymbol{J(D;\theta) = \alpha J_C(D;\theta) + (1-\alpha)\left(\gamma J_{F_G}(D;\theta) + (1-\gamma) J_{F_I}(D;\theta)\right) + \|\theta\|_2} \quad 11$$

where $\boldsymbol{J_C(D;\theta)}$ indicates the classification loss $\boldsymbol{J_{F_G}(D;\theta)}$ denotes the group loss and $\boldsymbol{J_{F_I}(D;\theta)}$ denotes the group loss and imposes individual fairness on the output of the model. This hybrid function has the following advantages.

- It can be implemented in any learning type.
- It can be implemented in machine learning model.
- Any group fairness metric and individual fairness metric can be used.
- By changing the parameters, the trade-off between fairness-accuracy and group fairness-individual fairness can be flexible according to the data set.

**3.1.2 Model architecture**. To implement our hybrid fairness method, we applied an architecture in Figure 11 on the COMPAS dataset.

In the data processing step, we have used COMPAS dataset to implement our model. According to ProPublica, this domain specific cleaning must be applied in order to get more accurate results:

- If the charge date of a defendants Compas scored crime was not within 30 days from when the person was arrested, we assume that because of data quality reasons, that we do not have the right offense.
- They have coded the recidivist flag -- is_recid -- to be -1 if we could not find a compas case at all.
- In a similar vein, ordinary traffic offenses -- those with a c_charge_degree of 'O' -- will not result in jail time are removed (only two of them).
- Dataset must be filtered the underlying data from Broward County to include only those rows representing people who had either recidivated in two years, or had at least two years outside of a correctional facility.

After this step, we processed the categorical and numerical data in the dataset as follows:

- We have applied SimpleImputer operation to numerical values for completing missing values.
- Using sensitive categorical data, we created a separate dataset to be used as a metric reference point.
- We converted the categorical data in the main data set into numerical data with one hot encoding.

The train/test split was determined as 0.6 for these two data sets. Logistic regression is used as a model for the classification problem. In the loss function, binary-crossentropy is used as classification loss error, demographic parity is used as group fairness loss and Theil Index is used as individual fairness loss.

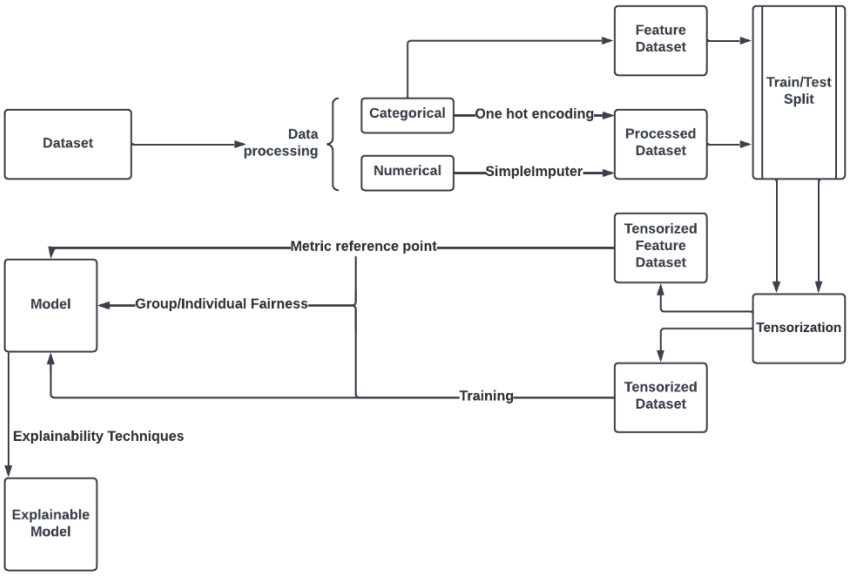


*Figure 11.* Model architecture diagram

**3.1.3 Logistic regression**. Logistic regression is a statistical analysis method for predicting a binary outcome such as yes or no based on previous observations of a data set. It is the appropriate regression analysis to be performed when the dependent variable is binary (binary). The logistic function is defined as in Equation 12:

$$\boldsymbol{logistic(\eta) = \frac{1}{1 + \exp{(\eta)}}} \quad 12$$

The simple linear regression formulation is in Equation 13.

$$\boldsymbol{y^{(\prime)} = \beta_0 + \beta_1 x_1^{(i)} + \cdots \beta_p x_p^{(i)}} \quad 13$$

Finally, the likelihood function is in Equation 14:

$$\boldsymbol{P\left(y^{(i)} = 1\right) = \frac{1}{1 + \exp{((-(\beta_0 + \beta_1 x_1^{(i)} + \cdots \beta_p x_p^{(i)})))}}} \quad 14$$

**3.1.4 Binary cross entropy loss**. Binary cross-entropy compares each of the predicted probabilities with the actual class output, which can be 0 or 1. The further the estimated probability is from the true label, the greater the cross-entropy loss. The defined error function is as in Equation 15.

$$L_{BCE} = -\frac{1}{n}\sum_{i=1}^{n} \hat{y}_i \log p(y_i) + (1-y)\log(1-p(y_i)) \quad (15)$$

**3.1.5 Demographic parity.** This fairness notion is also known as statistical parity in the literature and can be described as "a condition when a classifier produces outputs with equal probability for both protected and unprotected groups" (Dwork et al., 2012; Kusner et al., 2017). Equation of demographic parity can be shown in Equation 16.

$$P(P|A=0) = P(P|A=1) \quad (16)$$

Equation 16 states that the likelihood of an outcome should be equal regardless of group. The main disadvantage of this metric is that it is sufficient to positively label random individuals to provide the metric. Therefore, the success of the individual is not considered in this metric.

**3.1.6 Theil index.** The Theil index is a statistical measurement used to measure economic inequality over a population ("Theil Index," n.d.). Regional disparities are measured by a Theil entropy index, which is defined as in Equation 17.

$$T_T = T_{\alpha=1} = \frac{1}{N}\sum_{i=1}^{N} \frac{x_i}{\mu} \ln\left(\frac{x_i}{\mu}\right) \quad (17)$$

where μ is the mean income can be defined as in Equation 18:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \quad (18)$$

**3.1.7 Explainability.** Since we converted our training and test dataset into a tensor-based format, we used the Captum (Kokhlikyan et al., 2020) library with tensor-based explainability to interpret the logistic regression models we had trained. At Captum, we used Integrated Gradients, DeepLift, and GradientSHAP from so-called Primary Attribution methodologies whose purpose is to measure the impact of each feature on the model's output.

***3.1.7.1 Integrated gradients.*** Integrated gradients represent the integral of gradients can be efficiently approximated via a summation with respect to inputs along the path from a given baseline to input. One of the most important advantages of this method is that the method requires no modification to the original network and is extremely simple to implement. Equation 19 shows the formulation of the integrated gradient (Sundararajan, Taly, & Yan, 2017).

$$IntegratedGrads_i(x) ::= (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x_i))}{\partial x_i} d\alpha \quad (19)$$

***3.1.7.2 DeepLIFT.*** DeepLIFT is a back-propagation based approach, compares the activation of each neuron to its 'reference activation' and assigns contribution scores according to the difference (Shrikumar, Greenside, & Kundaje, 2017). DeepLIFT describes the difference in output from some 'reference' outputs in terms of the difference in input from some 'reference' inputs. DeepLIFT uses the multiplier to "blame" certain neurons for the difference in output. Multiplier can be explained as in Equation 20.

$$m_{\Delta t \Delta x} = \frac{C_{\Delta t \Delta x}}{\Delta x} \quad (20)$$

***3.1.7.3 GradientSHAP.*** GradientSHAP (Lundberg & Lee, 2017) is a gradient-based method to compute SHAP variables. Gradient SHAP adds Gaussian noise multiple times to each input sample, picks a random point along the path between the baseline and the input, and calculates the gradient of the outputs based on these random points selected. The final SHAP values represent the expected value of the gradients ("Algorithm Descriptions · Captum," n.d.).

## 3.2 Data Collection

In this section, we will evaluate the process from obtaining the data set required to carry out our experiment to its analysis.

**3.2.1 Data collection instruments.** The most used datasets we encountered in the literature review are as follows.

- COMPAS dataset
- German Credit Dataset
- Adult Census Dataset
- Communities and Crime Dataset
- Law School Dataset
- Student Performance Dataset

We used the COMPAS dataset for our study for some reasons stated below.

- The useful and redundant features in the COMPAS dataset are specified. Therefore, domain knowledge did not become very time-consuming in our work.
- The sociological impact of the COMPAS dataset is higher than the others. It is a serious problem for the society that there are protected features in determining a person's recidivism to commit crimes.
- The amount of data in the dataset is sufficient for prediction and the number of missing data is small.

The data set was also downloaded from https://github.com/propublica/compas-analysis and made ready for analysis with data engineering within the scope of the necessary instructions.

**3.2.2 Data collection procedures.** In this section we will provide information regarding the dataset.

***3.2.2.1 Data description.*** This dataset includes the prisoner's personal information, legal assessments, and estimates of recidivism. The dataset has 7214 rows and 53 columns in total, but these numbers will be reduced in data engineering part due to irrelevant columns, missing data and some filters. Table 14 shows the description of data and Table 15 depicts the first-five rows of the dataset.

Table 14

*Dataset Description of First Three and Last Three Columns*

| Type | id | age | juv_fel_count | end | event | two_year_recid |
|---|---|---|---|---|---|---|
| count | 7214.000 | 7214.000 | 7214.000 | 7214.000 | 7214.000 | 7214.000 |
| mean | 5501.256 | 34.818 | 0.067 | 553.437 | 0.383 | 0.451 |
| std | 3175.707 | 11.889 | 0.474 | 399.021 | 0.486 | 0.498 |
| mean | 1.000 | 18.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 25% | 2735.250 | 25.000 | 0.000 | 148.250 | 0.000 | 0.000 |
| 50% | 5509.500 | 31.000 | 0.000 | 530.500 | 0.000 | 0.000 |
| 75% | 8246.500 | 42.000 | 0.000 | 914.000 | 1.000 | 1.000 |
| max | 11001.000 | 96.000 | 20.000 | 1186.000 | 1.000 | 1.000 |

Table 15

*The First and Last Three Columns of First Five Rows of the Dataset*

| Number | id | name | first | end | event | two_year_recid |
|---|---|---|---|---|---|---|
| 0 | 1 | miguel hernandez | miguel | 327 | 0 | 0 |
| 1 | 3 | kevon dixon | kevon | 159 | 1 | 1 |
| 2 | 4 | ed philo | ed | 63 | 0 | 1 |
| 3 | 5 | marcu brown | marcu | 1174 | 0 | 0 |
| 4 | 6 | bouthy pierrelouis | bouthy | 1102 | 0 | 0 |

In the next section data engineering part will be explained.

***3.2.2.2 Data engineering.*** Data engineering is required for this dataset due to some reasons:

- Some records have empty values

- Some records have N/A values
- Some records are duplicated
- Some records are insufficient for training due to domain-specific reasons

According to ProPublica, this domain specific cleaning must be applied in order to get more accurate results:

- If the charge date of a defendants Compas scored crime was not within 30 days from when the person was arrested, we assume that because of data quality reasons, that we do not have the right offense.
- They have coded the recidivist flag -- is_recid -- to be -1 if we could not find a compas case at all.
- In a similar vein, ordinary traffic offenses -- those with a c_charge_degree of 'O' -- will not result in jail time are removed (only two of them).
- Dataset must be filtered the underlying data from Broward County to include only those rows representing people who had either recidivated in two years, or had at least two years outside of a correctional facility.

Furthermore, the number of columns should be reduced in this phase inasmuch as most of the columns are redundant for recidivism. We have taken account for these columns: age, c_charge_degree, race, age_cat, score_text, sex, priors_count, days_b_screening_arrest, decile_score, is_recid, two_year_recid, c_jail_in, c_jail_out. Table 16 shows number of null values in dataset.

Table 16

*Number of Empty Records for each Columns*

| Column | # of null values |
|---|---|
| age | 0 |
| c_charge_degree | 0 |
| race | 0 |
| age_cat | 0 |
| score_text | 0 |

Table 16 (continued)

| Column | # of null values |
|---|---|
| sex | 0 |
| priors_count | 0 |
| days_b_screening_arrest | 0 |
| decile_score | 0 |
| is_recid | 0 |
| two_year_recid | 0 |
| c_jail_in | 0 |
| c_jail_out | 0 |

And finally, Table 17 shows the first 5 rows of the modified dataset after the data engineering phase.

Table 17

*The First and Last Three Columns of First Five Rows of the Modified Dataset*

| Number | age | c_charge_degree | race | two_year_recid | c_jail_in | c_jail_out |
|---|---|---|---|---|---|---|
| 0 | 69 | F | Other | 0 | 2013-08-13 06:03:42 | 2013-08-14 05:41:20 |
| 1 | 34 | F | African-American | 1 | 2013-01-26 03:45:27 | 2013-02-05 05:36:53 |
| 2 | 24 | F | African-American | 1 | 2013-04-13 04:58:34 | 2013-04-14 07:02:04 |
| 3 | 44 | M | Other | 0 | 2013-11-30 04:50:18 | 2013-12-01 12:28:56 |
| 4 | 41 | F | Caucasian | 1 | 2014-02-18 05:08:24 | 2014-02-24 12:18:30 |

***3.2.2.3 COMPAS dataset.*** COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a case management and decision support tool developed and owned by Northpointe (now Equivant) used by U.S. courts to assess the likelihood of a defendant becoming a recidivist (“A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased against Blacks. It’s Actually Not

That Clear. - The Washington Post," n.d.; "Are Algorithms Building the New Infrastructure of Racism? - Nautilus | Science Connected," n.d.; "COMPAS (Software) - Wikipedia," n.d.; "Machine Bias — ProPublica," n.d.). COMPAS has been used by the U.S. states of New York, Wisconsin, California, Florida's Broward County, and other jurisdictions. The risk assessment consists of three scales namely:

- Pretrial release risk scale: Pretrial risk is a measure of an individual's potential not to appear and/or to commit new offenses during release.
- General recidivism scale: The general duplication scale is designed to predict new offenses after release and after the COMPAS rating has been given. The scale uses an individual's criminal history and partners, drug use, and signs of juvenile delinquency.
- Violent recidivism scale: The violent recidivism score is to predict violent crimes following release. The scale uses data or indicators that include a person's "history of violence, history of non-compliance, professional/educational issues, age of admission, and age of first arrest.

Table 18 illustrates the general information of data source and data content

Table 18

*COMPAS Recidivism Risk Score Information*

| Type | Information |
|---|---|
| Source | Broward County Clerk's Office, Broward County Sherrif's Office, Florida Department of Corrections, ProPublica |
| Data Released | March 2022 |
| Related Content | Machine Bias |
| Link | https://github.com/propublica/compas-analysis |
| File Content | 1 database, 6 csv files |

Csv files in the dataset are:

- compas-scores-raw.csv
- compas-scores-two-years-violent.csv
- compas-scores-two-years.csv
- compas-scores.csv
- cox-parsed.csv
- cox-violent-parsed.csv

Only one file (compas-scores-two-years.csv) are considered as a data source for our proposed method. We will describe the characteristics and analysis of the data we selected in the next section.

### 3.2.3 Data analysis procedures.

***3.2.3.1 Racial bias analysis.*** This section includes racial analysis of the dataset with race-score correlation, race distribution, true positive rate (TPR), true negative rate (TNR), false positive rate (FPR) and false negative rate (FNR).

In the first analysis which can be shown in Figure 12, racial distribution of the dataset has these values for each race is: i) African-American 3175 ii) Asian 31 iii) Caucasian 2103 iv) Hispanic 509 v) Native American 11 vi) Other 343. This shows us that the dataset is not racially evenly distributed, with Caucasians and African Americans making up the majority of the dataset. Under-represented features can lead to unfair decisions, notably in healthcare cases (Burlina et al., 2020; Puyol-Antón et al., 2021). The model will be developed considering this situation.

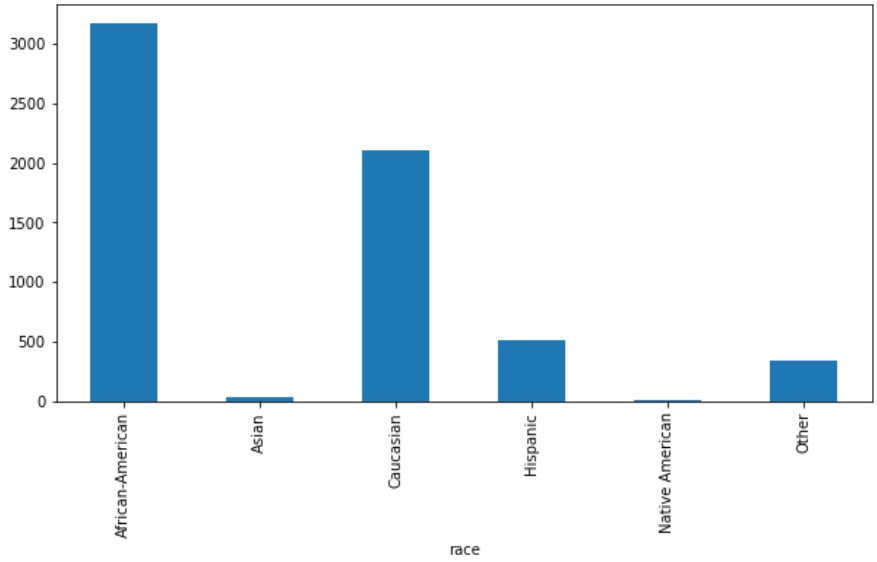


*Figure 12.* Racial distribution of the dataset

For the score distribution which is categorized as high low and medium, Figure 13 gives information about that low has the highest categorized score (3421) compared to medium (1607) and high (1144).

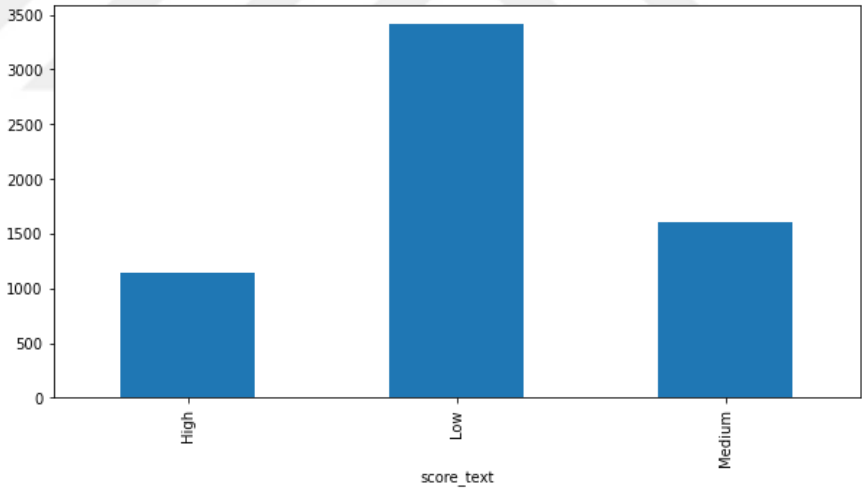


*Figure 13.* Score distribution of the dataset

As a next step, we have investigated the score distribution in for each race in the dataset. When the score distribution within the Caucasian and African-American races was compared, it was observed that Caucasians were marked higher at a considerably lower rate than African-Americans. Figure 14 shows the distribution graphically, and Table 19 shows exact values for each race.

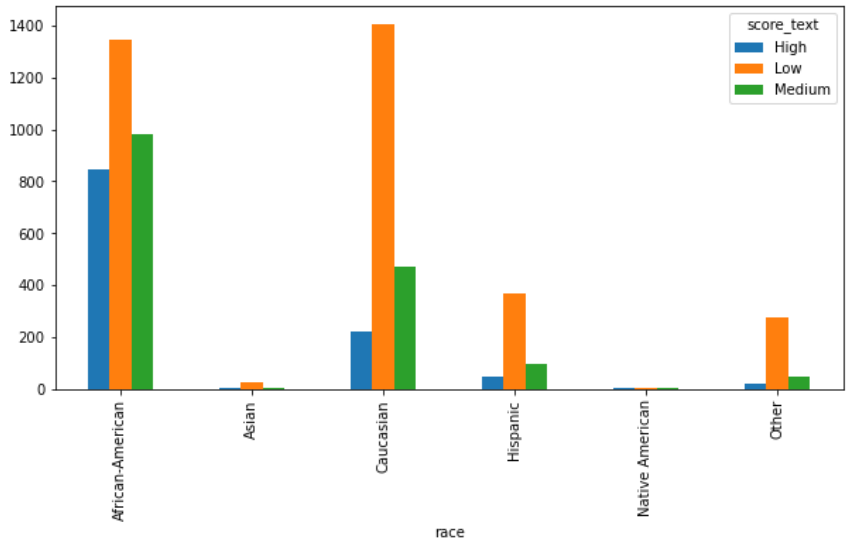


*Figure 14.* Score distribution for each race in dataset

Table 19

*Exact Values of Categorical Score Distribution for each Race*

| Race | Score Text | Count | Percentage |
|---|---|---|---|
| African-American | High | 845 | 26.614 |
| | Low | 1346 | 42.394 |
| | Medium | 984 | 30.992 |
| Asian | High | 3 | 9.667 |
| | Low | 24 | 77.419 |
| | Medium | 4 | 12.903 |
| Caucasian | High | 223 | 10.604 |
| | Low | 1407 | 66.904 |
| | Medium | 473 | 22.492 |
| Hispanic | High | 47 | 9.234 |
| | Low | 368 | 72.229 |
| | Medium | 94 | 18.468 |
| Native American | High | 4 | 36.364 |
| | Low | 3 | 27.273 |
| | Medium | 4 | 36.363 |
| Other | High | 223 | 6.414 |
| | Low | 1407 | 79,592 |
| | Medium | 473 | 13,994 |

Caucasians have marked as low more than 14% than African-Americans and African-Americans are marked as high higher than 16% than Caucasians. Comparing the percentage differences may not give accurate results, since the difference between the distribution numbers of other races and Caucasians is too large. These results shows that there might be racial bias in this dataset while percentages are taken account but further investigation is required.

For two-year recidivism (labeled as yes or no are labeled as 1 and 0 respectively) score, Table 20 points that there is a slight difference (around 3% 7% and 3%) between Caucasians and African-Americans. Figure 15 and Figure 16 shows the confusion matrix of recidivism prediction and score text for African-American and Caucasian respectively. Since the distance to both sides is equal, the mid scores are excluded in these matrices.

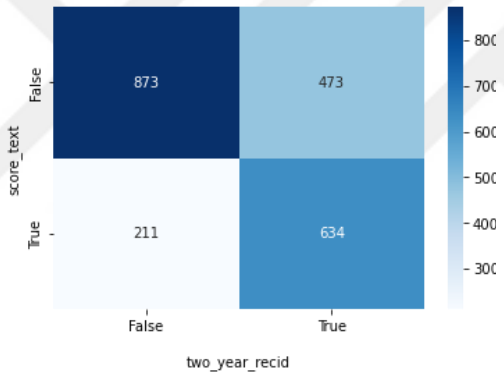


*Figure 15.* Confusion matrix for African-American

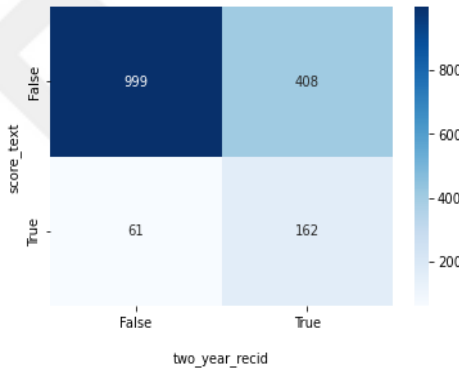


*Figure 16.* Confusion matrix for Caucasians

Table 20

*Exact Proportions of Score Distribution for each Race*

| Race | Score Text | Recid | No-Recid |
|---|---|---|---|
| African-American | High | 24.97 | 75.03 |
| | Low | 64.859 | 35.141 |
| | Medium | 43.700 | 56.300 |

Table 20 (continued)

| Race | Score Text | Recid | No-Recid |
|---|---|---|---|
| Asian | High | 33,333 | 66.667 |
| | Low | 87.5 | 12.5 |
| | Medium | 25.0 | 75.0 |
| Caucasian | High | 27,354 | 72.646 |
| | Low | 71.002 | 28.998 |
| | Medium | 46.723 | 53.277 |
| Hispanic | High | 42.553 | 57.447 |
| | Low | 70.109 | 29.891 |
| | Medium | 44.681 | 55.319 |
| Native American | High | 25.0 | 75.0 |
| | Low | 100.0 | 0.0 |
| | Medium | 50.0 | 50.0 |
| Other | High | 13.636 | 86.364 |
| | Low | 69.963 | 30.036 |
| | Medium | 52.083 | 47.916 |

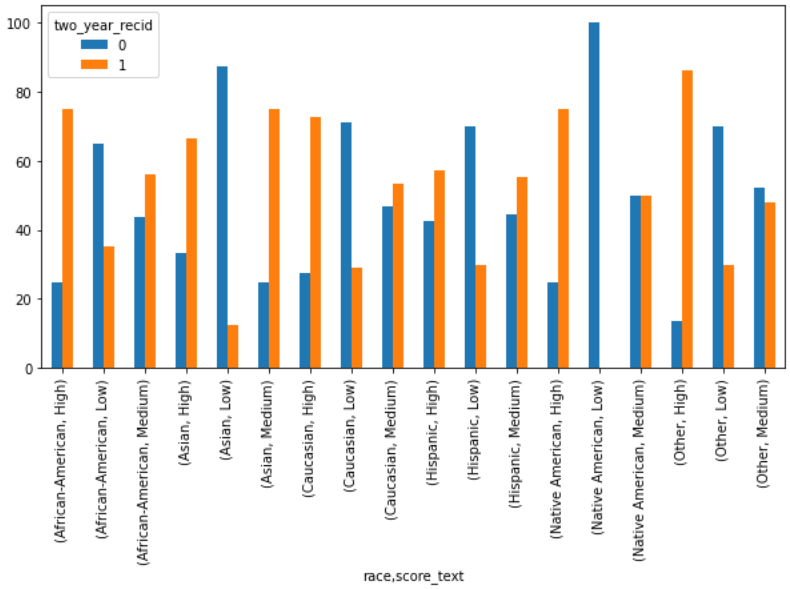


*Figure 17.* Score distribution for each race in dataset

Lastly for decile scores which are the numbering system for score categories (0-10), while African-American has equal distribution over the score, Caucasian has decreasing trend from 0 to 10. This clearly shows that Caucasians are marked with less decile scores than African-Americans. Figure 18 and Figure 19 show these distributions in graphic form and Table 21 shows for all races in tabular form.

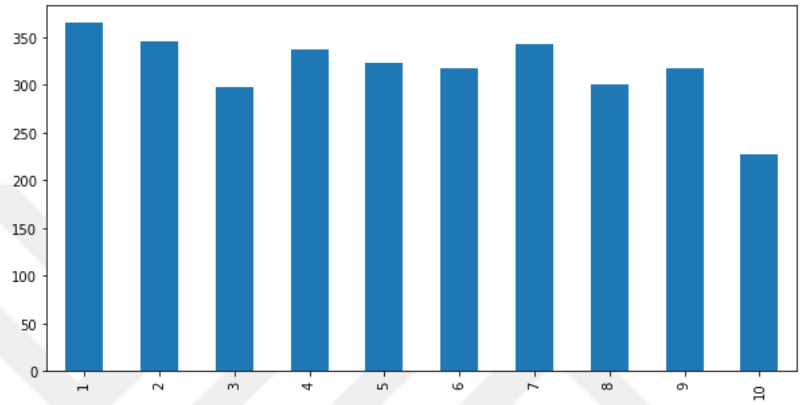


*Figure 18.* Decile point distribution in African-American

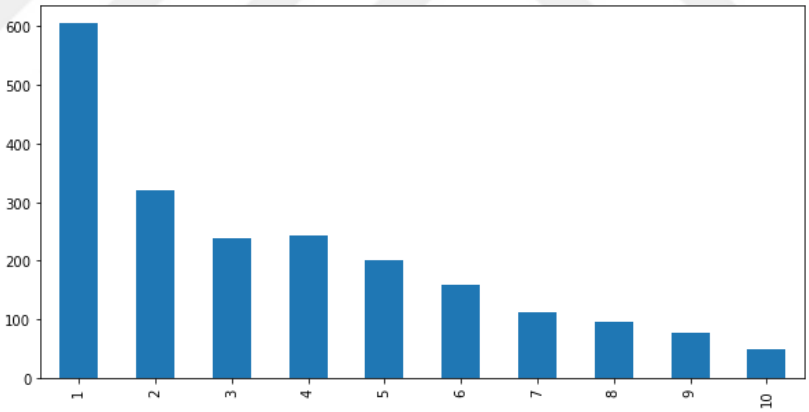


*Figure 19.* Decile point distribution in Caucasian

Table 21

*Exact Values of Score Distribution for each Race*

| Decile Score | Race | Count |
|---|---|---|
| 1 | African-American | 365 |
| | Asian | 15 |
| | Caucasian | 605 |

Table 21 (continued)

| Decile Score | Race | Count |
|---|---|---|
| | Hispanic | 159 |
| | Other | 142 |
| 2 | African-American | 346 |
| | Asian | 4 |
| | Caucasian | 321 |
| | Hispanic | 89 |
| | Native American | 2 |
| | Other | 60 |
| 3 | African-American | 298 |
| | Asian | 5 |
| | Caucasian | 238 |
| | Hispanic | 73 |
| | Native American | 1 |
| | Other | 32 |
| 4 | African-American | 337 |
| | Caucasian | 243 |
| | Hispanic | 47 |
| | Other | 39 |
| 5 | African-American | 323 |
| | Asian | 1 |
| | Caucasian | 200 |
| | Hispanic | 47 |
| | Other | 39 |
| 6 | African-American | 318 |
| | Asian | 2 |
| | Caucasian | 160 |
| | Hispanic | 27 |
| | Native American | 1 |
| | Other | 20 |
| 7 | African-American | 343 |
| | Asian | 1 |

Table 21 (continued)

| Decile Score | Race | Count |
|---|---|---|
| | Caucasian | 113 |
| | Hispanic | 28 |
| | Native American | 2 |
| | Other | 9 |
| 8 | African-American | 301 |
| | Asian | 2 |
| | Caucasian | 96 |
| | Hispanic | 14 |
| | Other | 7 |
| 9 | African-American | 227 |
| | Asian | 1 |
| | Caucasian | 50 |
| | Hispanic | 16 |
| | Native American | 2 |
| | Other | 7 |
| 10 | African-American | 227 |
| | Asian | 1 |
| | Caucasian | 50 |
| | Hispanic | 16 |
| | Native American | 2 |
| | Other | 8 |

We can clearly state that decile scoring is biased by racial bias which is proved by data analysis and exploration.

***3.2.3.2 Age bias analysis.*** In this section we will present some correlation analysis between age and scores. For correlation analysis:

- -0.4037 for age-decile score correlation
- -0.1891 for age-two_year recidivism correlation

To show bias in age sub-groups, further investigation is required therefore we visualize subgroups in decile score and recidivism cases. Figure 20 shows that middle aged defendants get lower score than people who is younger. Likewise, Figure 21 depicts that middle-aged defendant has marked no recidivism score more than people who is under younger. The gender bias as well as race and age bias are proven in the *Generalized linear model for multi bias analysis.* using the generalized linear model.

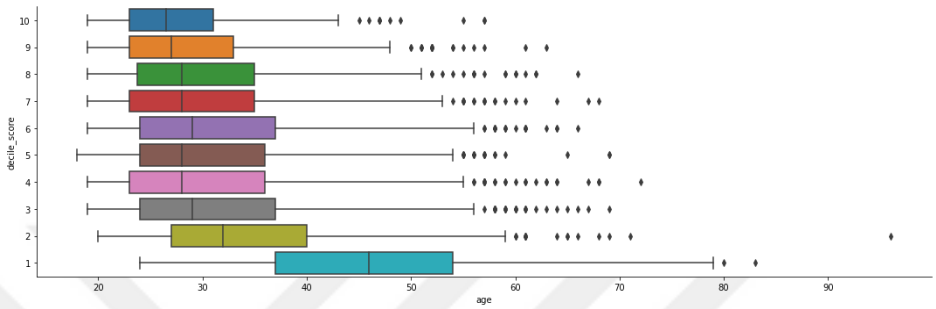


*Figure 20.* Decile point distribution in age sub-groups

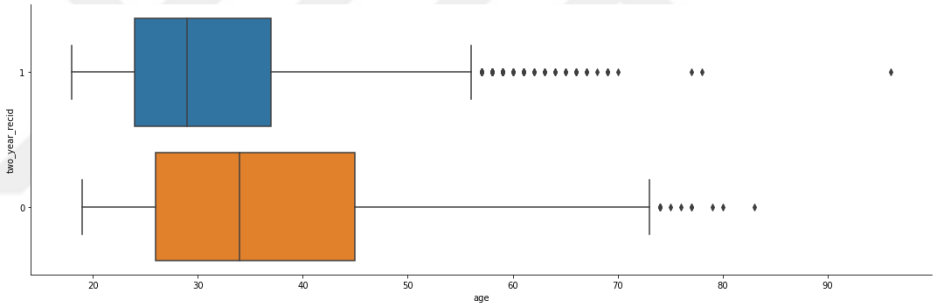


*Figure 21.* Recidivism point distribution in age sub-groups

***3.2.3.3 Generalized linear model for multi bias analysis.*** To prove our claims, we have implemented Generalized Linear Model (GLM) from statsmodel in R. We have selected R instead of Python for this analysis because, in Python, categorical variables should be encoded and R does intersection of variables itself. And also, there is output differences between R and Python and R looks more stable than Python in GLM. We have selected binomial distribution as a family and we have implemented this formula in Equation 21.

$$\boldsymbol{E_{Feature} = F_{Score} \sim F_{Gender} + F_{Age} + F_{Race} + P_{Priors} + F_{Crime} + R_{two\ year}} \quad 21$$

We have selected $F_{Score}$as endog and other variables as exog. Afterwise, we have obtained these results which can be found in Table 22.

Table 22

*Values Calculated by GLM*

| Feature | Estimate | Std. Error | z value |
|---|---|---|---|
| (Intercept) | -1.525 | 0.078 | -19.430 |
| gender_factorFemale | 0.221 | 0.079 | 2.783 |
| age_factorGreater than 45 | -1.355 | 0.099 | -13.682 |
| age_factorLess than 25 | 1.308 | 0.075 | 17.232 |
| race_factorAfrican-American | 0.477 | 0.069 | 6.881 |
| race_factorAsian | -0.254 | 0.478 | -0.532 |
| race_factorHispanic | -0.428 | 0.128 | -3.344 |
| race_factorNative American | 1.394 | 0.766 | 1.820 |
| race_factorOther | -0.826 | 0.162 | -5.098 |
| priors_count | 0.268 | 0.011 | 24.221 |
| crime_factorM | -0.311 | 0.066 | -4.677 |
| two_year_recid | 0.685 | 0.064 | 10.713 |

We can extract bias effect from the GLM with the help of binomial distribution. For control value we have calculated as $c = \exp(-1.525)\ /(1 + \exp\ (-1.525))$ and all calculations to discover bias are used on this control term.

- Racial bias: African Americans get higher scores than Caucasians with 45% rate (Proof: $\exp(0.472)\ /(1 - c + (c * \exp\ (0.477)))$
- Gender bias: Women get higher scores than men with 19% rate (Proof: $\exp(0.221)\ /(1 - c + (c * \exp(0.221)))$
- Age bias: People under 25 get higher scores than middle aged defendants with 250% rate (Proof: $\exp(1.308)\ /(1 - c + (c * \exp\ (1.308)))$

**3.2.4 Reliability and validity.** Validity and reliability are the two most important considerations for the trustworthiness of the study. We will examine Validity both internally and externally.

***3.2.4.1 Internal validity***. For the internal validity, this data set includes external variable independent constant outputs and the constant features used for this output, instead of containing variables that will affect the work in certain period intervals.

***3.2.4.2 External validity***. For the external validity, although this dataset is a dataset containing recidivism, it is one of the most used datasets in the field of fair artificial intelligence and since it is a dataset containing protected features, it has the capacity to be generalized for other decision-making problems.

***3.2.4.3 Reliability*** Data reliability is directly proportional to the reliability of the company that created the data. ProPublica is a non-profit organization that conducts research activities on issues of public interest. In return for these efforts, it became the first online journalism platform to win the Pulitzer Prize. The COMPAS software was created by a company called Northpointe, Inc, and this raises doubts about the data. However, the fact that our download source is ProPublica and it is the most used dataset in the field of fair artificial intelligence, and there is no discussion about its reliability at the same time shows that the dataset is reliable.

## 3.3 Limitations

In this section we will investigate some of the limitations of our work.

**3.3.1 Fairness definitions.** Our approach is to bring together individual and group fairness approaches, minimizing the disadvantages created by both fairness metrics at certain points and presenting this with the output of explainability. In this methodology, there are certain limitations related to both the ontological structure of artificial intelligence, which we mentioned in the fair systematic literature review, and our problem space. A limitation in our study is which fairness metrics should be used in examples such as datasets where there are more than one intersecting conserved features or where the positive and negative distributions contain an imbalance on the

basis of groups. For this, the definitions of fairness metrics and the data set feature vectors should be examined together.

**3.3.2 Fairness interaction.** Since our study proposes a fairness loss function as a combination of group and individual fairness metrics, it is a limitation on our side whether which individual fairness function is compatible with which group fairness function. Since minimizing the error function that will occur with two completely inversely proportional fairness metrics will require great effort, the interaction between group and individual fairness metrics should be examined.

**3.3.3 Feature ranking.** One of the limitations is on which feature a fair model will be created in cases where there is more than one protected feature. Since if a dataset includes indirectly related features (such as Race-Country), bias in these cases constitutes a limitation for our methodology. As a solution to this limitation, methodologies such as protected feature ranking should be studied and developed.

**3.3.4 Generalization.** The applicability problem to real-time systems, which is one of the structural problems of fair artificial intelligence, is also valid for our proposed methodology. How our model can be applied in real-time systems or how it will behave in a different data set has been seen as a limitation by us. In addition, we are wondering how the trade-off between individual-group fairness and accuracy-fairness will be evaluated in the domain adaptation part.

# Chapter 4: Findings

In this section, the experimental parameters, necessary packages and outputs used for the hypothesis are presented.

## 4.1 Required Packages

In this study, we used the following libraries and frameworks for the following reasons:

- AIF360: Generalized Entropy Error implementation
- Captum: Tensor-based explainability modules
- Fairlearn: Implementing group fairness metrics, dashboard
- Matplotlib: Visualization of experimental results
- Pytorch: Logistic regression implementation, gradient calculation
- Sklearn: Data engineering operations and train/test splitting
- Tensorflow: Tensor operations
- TQDM: Machine learning progress

## 4.2 Experimental Results

Experiment was conducted with NVIDIA RTX 2060 Graphics Card and 32GB DDR4 ram with the mentioned libraries and frameworks installed. The logistic regression model was trained using stochastic gradient descent with 10000 epochs and 0.01 learning rate. 25 models were trained using values of 1, 0.5, 0.05, 0.005, 0 for both alpha and gamma parameters. Table 23 shows all the evaluation outputs of the 25 models created with these parameters. As expected, the full accuracy-based model directly reflects the bias in the dataset, while interestingly, the full-fairness-based models contain a low accuracy rate, but at the same time a low error value. This is due to the fact that the concept of fairness only looks at group-based selection variation, but does not consider the difference between input and output. This causes almost all of the outputs in the model to be labeled with a single value, because if all data are marked the same as positive or negative, there is no effect of group-based difference.

Table 23

*Detailed Evaluation Table*

| Model | Alpha | Gamma | Epoch | Train Acc | Test Acc | Train Loss | Test Loss | Train GF Loss | Test GF Loss | Train IF Loss | Test IF Loss |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LR | 1.0 | 1.0 | 10000 | 0.976 | 0.968 | 0.118 | 0.141 | 0.428 | 0.292 | 0.236 | 0.014 |
| LR | 1.0 | 0.5 | 10000 | 0.976 | 0.968 | 0.119 | 0.141 | 0.428 | 0.292 | 0.236 | 0.014 |
| LR | 1.0 | 0.05 | 10000 | 0.976 | 0.968 | 0.118 | 0.140 | 0.428 | 0.292 | 0.236 | 0.014 |
| LR | 1.0 | 0.005 | 10000 | 0.976 | 0.968 | 0.118 | 0.140 | 0.428 | 0.292 | 0.236 | 0.014 |
| LR | 1.0 | 0.0 | 10000 | 0.976 | 0.968 | 0.118 | 0.140 | 0.428 | 0.292 | 0.236 | 0.014 |
| LR | 0.5 | 1.0 | 10000 | 0.976 | 0.967 | 0.293 | 0.234 | 0.428 | 0.292 | 0.236 | 0.014 |
| LR | 0.5 | 0.5 | 10000 | 0.976 | 0.967 | 0.188 | 0.164 | 0.428 | 0.292 | 0.236 | 0.014 |
| LR | 0.5 | 0.05 | 10000 | 0.976 | 0.967 | 0.093 | 0.101 | 0.428 | 0.292 | 0.236 | 0.014 |
| LR | 0.5 | 0.005 | 10000 | 0.976 | 0.968 | 0.083 | 0.093 | 0.428 | 0.292 | 0.236 | 0.014 |
| LR | 0.5 | 0.0 | 10000 | 0.976 | 0.968 | 0.082 | 0.093 | 0.428 | 0.292 | 0.236 | 0.014 |
| LR | 0.05 | 1.0 | 10000 | 0.819 | 0.808 | 0.373 | 0.163 | 0.369 | 0.147 | 0.268 | 0.107 |
| LR | 0.05 | 0.5 | 10000 | 0.830 | 0.820 | 0.214 | 0.140 | 0.311 | 0.147 | 0.264 | 0.098 |
| **LR** | **0.05** | **0.05** | **10000** | **0.851** | **0.835** | **0.107** | **0.107** | **0.303** | **0.107** | **0.263** | **0.088** |
| **LR** | **0.05** | **0.005** | **10000** | **0.825** | **0.811** | **0.114** | **0.122** | **0.327** | **0.111** | **0.267** | **0.103** |
| **LR** | **0.05** | **0.0** | **10000** | **0.862** | **0.846** | **0.090** | **0.099** | **0.303** | **0.073** | **0.260** | **0.081** |
| LR | 0.005 | 1.0 | 10000 | 0.678 | 0.664 | 0.337 | 0.170 | 0.336 | 0.168 | 0.276 | 0.201 |
| LR | 0.005 | 0.5 | 10000 | 0.733 | 0.706 | 0.193 | 0.327 | 0.227 | 0.470 | 0.287 | 0.181 |
| LR | 0.005 | 0.05 | 10000 | 0.708 | 0.682 | 0.184 | 0.209 | 0.276 | 0.361 | 0.297 | 0.199 |
| LR | 0.005 | 0.005 | 10000 | 0.714 | 0.702 | 0.183 | 0.197 | 0.268 | 0.321 | 0.311 | 0.195 |
| LR | 0.005 | 0.0 | 10000 | 0.727 | 0.708 | 0.163 | 0.183 | 0.255 | 0.229 | 0.290 | 0.181 |
| LR | 0.0 | 1.0 | 10000 | 0.438 | 0.461 | 0.002 | 0.005 | 0.002 | 0.005 | 0.051 | 0.054 |
| LR | 0.0 | 0.5 | 10000 | 0.560 | 0.534 | 0.195 | 0.217 | 0.0 | 0.0 | 0.391 | 0.434 |
| LR | 0.0 | 0.05 | 10000 | 0.561 | 0.535 | 0.371 | 0.412 | 0.001 | 0.001 | 0.392 | 0.434 |
| LR | 0.0 | 0.005 | 10000 | 0.562 | 0.539 | 0.386 | 0.424 | 0.002 | 0.005 | 0.392 | 0.426 |

According to the parameters in the table, we can classify the models as follows:

$$Model = \begin{cases} Full\ accuracy\ focused & if\ \alpha = 1 \\ Semi - fair\ full\ only\ group & if\ \alpha = 0.5\ and\ \gamma = 1 \\ Semi\ fair\ equal\ individual - group & if\ \alpha = 0.5\ and\ \gamma = 0.5 \\ Semi - fair\ full\ only\ individual & if\ \alpha = 0.5\ and\ \gamma = 0 \\ Full\ fairness\ focused & if\ \alpha = 0 \end{cases}$$

Models other than these values can be considered as metric weighted models. As can be seen from the Table 23, the most optimal models are those in which fairness is intense and individual fairness is high in this fairness.

### 4.3 Explainability

We used captum, a tensor-based explainability library, for explainability, and in this section, we will present the explainability of the models we have chosen as promising. The following methods were used for explainability.

- Integrated gradients: Integrated Gradient is an interpretability or explainability technique for deep neural networks that visualizes the input feature significance that contributes to the model's prediction.
- DeepLIFT (Deep Learning Important FeaTures) a novel algorithm to assign importance score to the inputs for a given output (Shrikumar et al., 2017)
- GradientShap (Lundberg & Lee, 2017) approximates SHAP values by computing the expectations of gradients by randomly sampling.

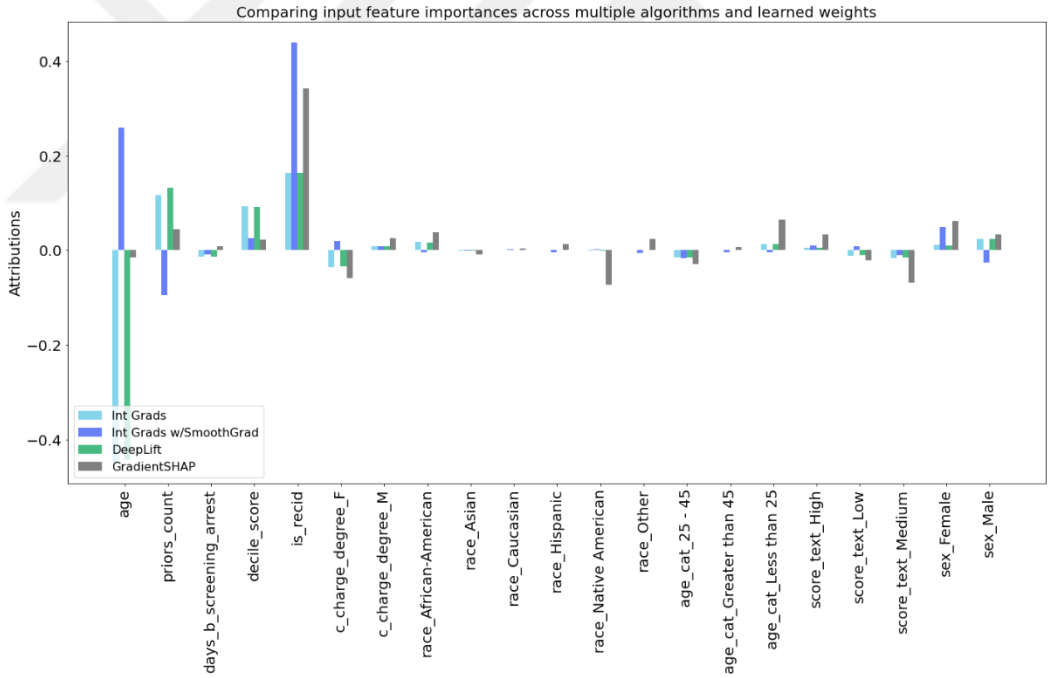


*Figure 22.* Explainability results for α= 0.05 and γ= 0

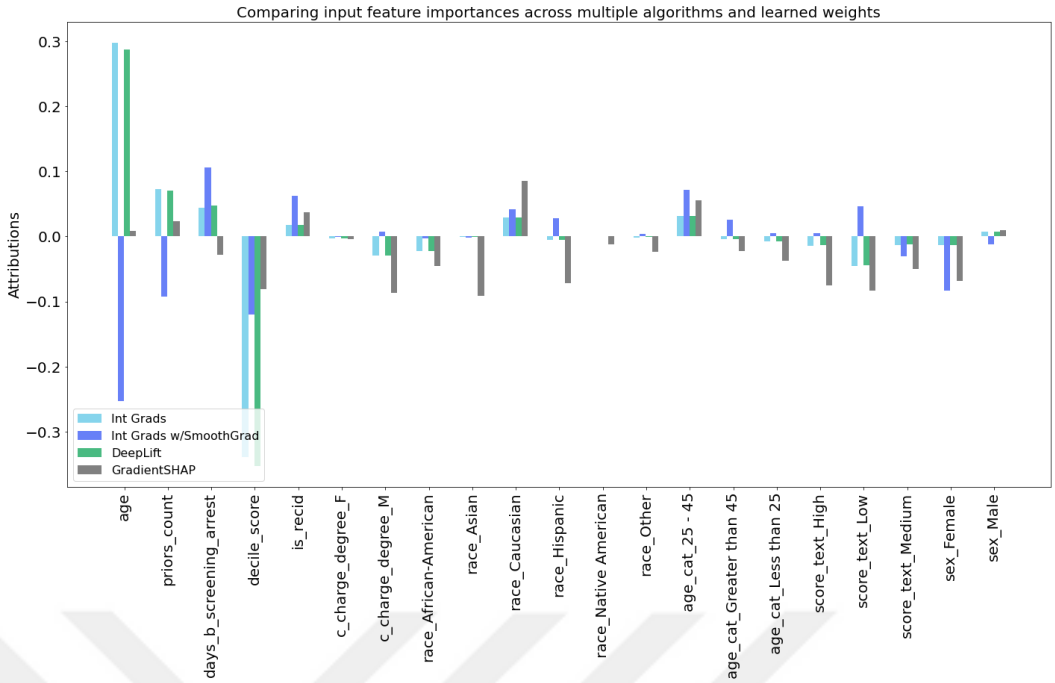


*Figure 23.* Explainability results for α= 0.05 and γ= 0.05

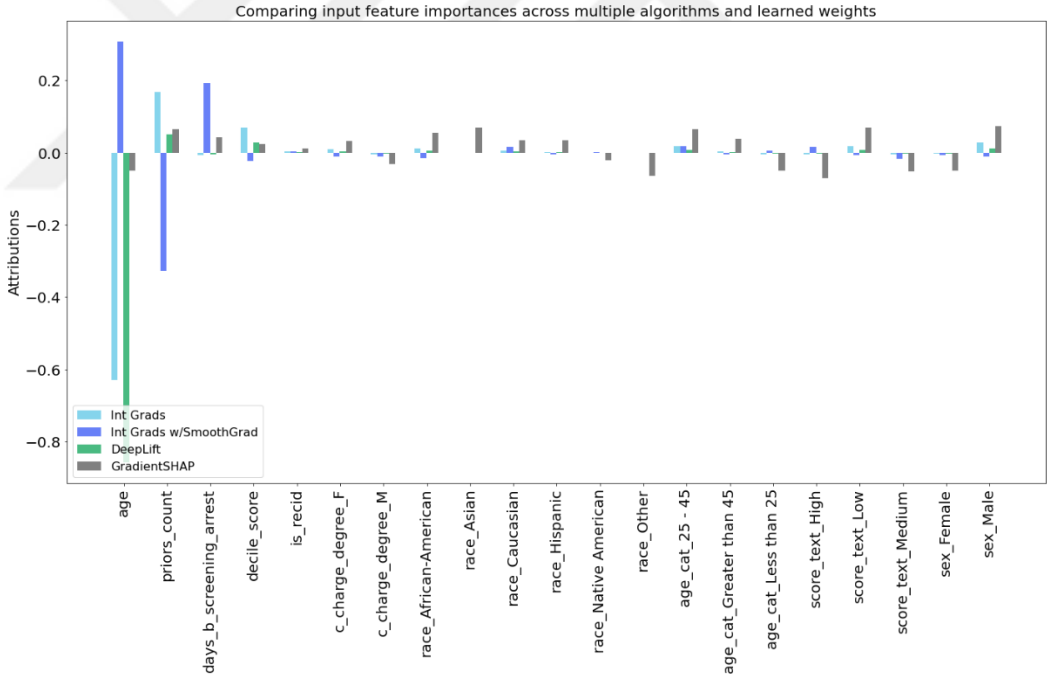


*Figure 24.* Explainability results for α= 0.05 and γ= 0.005

According to the explainability output, African-Americans, who are the biggest disadvantaged group for α= 0.05, and γ= 0.005, appear to have positive feature scores. For the other two (0.05,0) varying γ values, the trait effect with the same race includes very small scores close to 0. At the value of γ= 0.005, it can be seen that the effect of different trait vectors belonging to gender is very close to each other. This shows us

that properly selected group and individual fairness weights give more optimal results on an explicable basis. It can be said that when $\gamma$ values are high, feature effect is protective attribute-based sharper, while protective attribute-based features are softer at low values. This shows that explainability presents the effect effectively.

# Chapter 5: Discussions and Conclusions

This section includes an overview of the outcomes, ethical impact, summary and motivations for future work.

## 5.1 Discussion on Findings for Research Question

In this study, we defined a hybrid fairness metric to reduce the disadvantage of group fairness, which is a frequently used metric in fair artificial intelligence methodologies, on individuals. We used logistic models we trained with different $\alpha$ and $\gamma$ parameters to evaluate this proposed metric. As a result, it was seen that the individual fairness metric can also affect group fairness positively at certain $\gamma$ values, and that individual fairness can make a positive contribution to accuracy. Also, accuracy-only models may include bias in the dataset, and fairness-only models may result in labeling all data identically, which can be termed "all or none". Within the scope of explainability, while the protective attribute effect is softer in individual fairness-based models, it is sharper in group-based fairness-based models. This shows us that group-based models of fairness achieve this by sharpening and trying to equalize the effect of protective attributes.

## 5.2 Ethical Impact

With the progress in artificial intelligence, systems based on artificial intelligence have started to be used frequently in daily life. However, in decision-critical systems such as recruitment, scholarship and loan applications, prejudices based on the history cause artificial intelligence systems to carry these prejudices. The models proposed to prevent this are aimed at eliminating ins towards privileged groups, which are largely comprised of individuals with disadvantaged attributes. However, group-based fairness is aimed at equalizing group-based ratios, and it may have effects such as unsuccessful individuals gaining an advantageous reward or successful individuals being disadvantaged. To avoid this problem, we defined fairness as the problem of finding the optimum ratio for the selected group and individual fairness metrics.

One of the major social problems associated with the decision-making system is its use in the job market. Equal Employment Opportunity Commission (EEOC) states that factors such as particular race, color, religion, sex (including pregnancy, sexual orientation, or gender identity), national origin, disability status or age (40 or older) may have a negative impact on recruitment ("Hiring Practices That Have a Negative Effect on Certain Applicants | U.S. Equal Employment Opportunity Commission," n.d.). Likewise in the United States, anti-discrimination law is set in Title VII of the Civil Rights Act of 1964, which limits the types of conduct that employers can engage in. The U.S. Supreme Court, Griggs v. Duke Power Co. ("Griggs v. Duke Power Co. :: 401 U.S. 424 (1971) :: Justia US Supreme Court Center," n.d.) ruled those certain behaviors – even unintentional and illegal – that could lead to discriminatory consequences – were the doctrine of "different influence". The Court has not based and statistically defined disparate impact, but the Equal Employment Opportunity Commission (EEOC) decides on disparate impact based on the 80% selection rate rule ("Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures | U.S. Equal Employment Opportunity Commission," n.d.). The definition of the rule is " A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact." ("29 CFR § 1607.4 - Information on Impact. | CFR | US Law | LII / Legal Information Institute," n.d.). While our definition of fairness includes the 80% rule, it also includes the concepts of individual fairness advocating individual equality.

In our study, instead of criticizing the existing system, we plan to present the problems of the existing system and create a new perspective. From this point of view, our assumption that fair artificial intelligence studies proposed to prevent unfair decisions made to groups may cause individual-based problems, we hope that instead of reducing the effect of group-based fairness, the fairness of the approaches made to the individuals forming the groups and the groups they belong to will be evaluated together, and we hope that the concept of fairness will be evaluated on this level.

## 5.3 Conclusions

The concept of fairness has become as important a metric as accuracy for artificial intelligence methodologies with the increasing use of artificial intelligence in real world systems. The biases contained in historical data have led to the proposal of fair artificial intelligence methodologies in the field of artificial intelligence, which can be categorized as reducing the bias of these data, training the bias of the models, and making the model outputs unbiased. Fair AI models of these categories mostly focus on group fairness, which aims to ensure that groups grouped according to protected attributes have equal statistical output rates. However, in group fairness, the situation of individuals to get positive or negative results in an unfair way may contradict the individual fairness aiming at individual equality. We defined the concept of fairness as the problem of finding the optimum point of individual and group-based fairness evaluations in artificial intelligence, and as a result, we developed the methodology called hybrid fairness.

In the proposed function in Methodology section, we control the trade-off between accuracy and fairness with the $\alpha$ parameter, and the trade-off between the selected group and individual fairness with the $\gamma$ parameter. We conducted the experiments on logistic regression using demographic parity for group fairness and Theil index as individual fairness metric on COMPAS dataset. As a result of our experiments, we discovered that individual fairness also contributes positively to group fairness at some points.

In addition, we have presented the logistic regression models, which we trained the most promising values using gradient-based attribution methods, as explainable, since we believe that fairness is a metric that should also be presented with explainability, rather than just a notion based on minimizing a statistical error function.

We planned to examine the hybrid fairness we proposed, the effect of multiple group fairness metrics and individual metrics, and examine the datasets with excessive group imbalance, as a future study with the equations in the Recommendations ection.

We mentioned that hybrid fairness is flexible in terms of group and individual fairness metrics, but due to the definition of fairness metrics, determining which group-individual fairness metric pair is the most efficient and which is the most inefficient is a limitation for us at the moment. In addition, generalizability, which is unfortunately a general problem of fair artificial intelligence models, can be considered a constraint for us.

Finally, our study considers fairness as an optimization problem between the group and individual fairness, as we mentioned. Our aim here is not to criticize models that focus solely on group or individual fairness but to offer a different perspective to future studies. Likewise, our proposition of fairness with explainability is because we want to show how important explainability is to fairness, not to diminish the significance of its statistical metric.

### 5.4 Recommendations

First of all, a methodology that can be defined as the combination of more than one group fairness or more than one individual fairness metric can be counted as a constraint due to the impossibility of providing more than one metric in fair artificial intelligence (Chouldechova, 2017). However, it is planned as a future study how the use of multiple non-contradictory fairness metrics will have an effect on fair artificial intelligence. This conditionally extended hybrid fairness function can be written as in Equation 22:

$$\begin{aligned} \boldsymbol{J(D;\theta) = \alpha J_C(D;\theta)} \\ &\boldsymbol{+ (1-\alpha)\big(\gamma(R_{G1}J_{F_{G1}}(D;\theta)} \\ &\boldsymbol{+\ R_{G2}J_{F_{G2}}(D;\theta) \dots R_{GN}J_{F_{GN}}(D;\theta)) + (1-\gamma)(R_{I1}J_{I_I}(D;\theta)} \\ &\boldsymbol{+ R_{I2}J_{I_2}(D;\theta) \dots R_{IN}J_{I_N}(D;\theta))\big) + \beta\|\theta\|_2} \end{aligned} \qquad 22$$

Another future work may also apply to datasets with very uneven group weights. In that case, a conversion like this can be used in Equation 23.

$$\boldsymbol{J(D;\theta) = \alpha J_C(D;\theta) + (1-\alpha)\big(J_F(D;\theta)\big) + \beta\|\theta\|_2} \quad 23$$

$$\boldsymbol{J_F = \sum_{i=1}^{\#\ of\ group} \left( R_G G_F(i) + (1 - R_G \sum_{j=1}^{N} I_F(j)) \right)} \quad 24$$

$$\boldsymbol{R_G = \frac{|G|}{|G_{Biggest}|}} \quad 25$$

Where $|G|$ denotes the population size and $R_G$ represents the strength of the group. In this methodology, an increase in population size increases the group effect, and a decrease in population size increases the individual fairness effect. Since this method has flexible architecture, γ parameter is not needed.

## REFERENCES


– European Union Agency for Fundamental Rights, F. (n.d.). *Data quality and artificial intelligence-mitigating bias and error to protect fundamental rights HELPING TO MAKE FUNDAMENTAL RIGHTS A REALITY FOR EVERYONE IN THE EUROPEAN UNION FRA Focus Contents*. https://doi.org/10.2811/615718

4 possible ways to avoid big data bias | European Union Agency for Fundamental Rights. (n.d.). Retrieved May 14, 2022, from https://fra.europa.eu/en/news/2018/4-possible-ways-avoid-big-data-bias

29 CFR § 1607.4 - Information on impact. | CFR | US Law | LII / Legal Information Institute. (n.d.). Retrieved May 15, 2022, from https://www.law.cornell.edu/cfr/text/29/1607.4

A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. - The Washington Post. (n.d.). Retrieved May 14, 2022, from https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/

Addressing Algorithmic Discrimination in the European Union - A Path For Europe (PfEU). (n.d.). Retrieved May 14, 2022, from https://pathforeurope.eu/addressing-algorithmic-discrimination-in-the-european-union/

Adel, T., Valera, I., Ghahramani, Z., & Weller, A. (2019). One-network adversarial fairness. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*. https://doi.org/10.1609/aaai.v33i01.33012412

Agarwal, A., Beygelzimer, A., Dudfk, M., Langford, J., & Hanna, W. (2018). A reductions approach to fair classification. *35th International Conference on Machine Learning, ICML 2018*, *1*.

Aivodji, U., Bidet, F., Gambs, S., Ngueveu, R. C., & Tapp, A. (2021). Local data debiasing for fairness based on generative adversarial training. *Algorithms*, *14*(3). https://doi.org/10.3390/a14030087

Algorithm Descriptions · Captum. (n.d.). Retrieved May 15, 2022, from https://captum.ai/docs/algorithms

Are Algorithms Building the New Infrastructure of Racism? - Nautilus | Science Connected. (n.d.). Retrieved May 14, 2022, from https://nautil.us/are-algorithms-building-the-new-infrastructure-of-racism-6874/

Awasthi, P., Beutel, A., Kleindessner, M., Morgenstern, J., & Wang, X. (2021). Evaluating fairness of machine learning models under uncertain and incomplete information. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3442188.3445884

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Müller, K. R. (2010). How to explain individual classification decisions. *Journal of Machine Learning Research*, *11*.

Balashankar, A., & Lees, A. (2019). *Fairness Sample Complexity and the Case for Human Intervention*. https://doi.org/10.1145/nnnnnnn

Barry-Jester, A. M., Casselman, B., & Goldstein, D. (2015). The New Science of Sentencing. *The Marshall Project*.

Bobadilla, J., Lara-Cabrera, R., González-Prieto, Á., & Ortega, F. (2021a). Deepfair: Deep learning for improving fairness in recommender systems. *International Journal of Interactive Multimedia and Artificial Intelligence*, *6*(6). https://doi.org/10.9781/ijimai.2020.11.001

Bobadilla, J., Lara-Cabrera, R., González-Prieto, Á., & Ortega, F. (2021b). Deepfair: Deep learning for improving fairness in recommender systems. *International Journal of Interactive Multimedia and Artificial Intelligence*, *6*(6). https://doi.org/10.9781/ijimai.2020.11.001

Borges, R., & Stefanidis, K. (2019). Enhancing long term fairness in recommendations with variational autoencoders. *11th International Conference on Management of Digital EcoSystems, MEDES 2019*. https://doi.org/10.1145/3297662.3365798

Brandão, M., Jirotka, M., Webb, H., & Luff, P. (2020). *Fair navigation planning: a resource for characterizing and designing fairness in mobile robots*.

Burlina, P., Paul, W., Mathew, P., Joshi, N., Pacheco, K. D., & Bressler, N. M. (2020). Low-Shot Deep Learning of Diabetic Retinopathy with Potential Applications to Address Artificial Intelligence Bias in Retinal Diagnostics and Rare Ophthalmic Diseases. *JAMA Ophthalmology*, *138*(10). https://doi.org/10.1001/jamaophthalmol.2020.3269

Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). Building classifiers with independency constraints. *ICDM Workshops 2009 - IEEE International Conference on Data Mining*. https://doi.org/10.1109/ICDMW.2009.83

Calders, T., Karim, A., Kamiran, F., Ali, W., & Zhang, X. (2013). Controlling attribute effect in linear regression. *Proceedings - IEEE International Conference on Data Mining, ICDM*. https://doi.org/10.1109/ICDM.2013.114

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334). https://doi.org/10.1126/science.aal4230

Cao, Y., Berend, D., Tolmach, P., Levy, M., Amit, G., Shabtai, A., … Liu, Y. (2020). Fairness matters – a data-driven framework towards fair and high performing facial recognition systems. *ArXiv*.

Celis, L. E., Keswani, V., & Vishnoi, N. K. (2020). *Data preprocessing to mitigate bias: A maximum entropy based approach*.

Chakraborti, T., Patra, A., & Noble, J. A. (2020). Contrastive Fairness in Machine Learning. *IEEE Letters of the Computer Society*, *3*(2). https://doi.org/10.1109/locs.2020.3007845

Chen, Jiahao, Kallus, N., Mao, X., Svacha, G., & Udell, M. (2019). Fairness under unawareness: Assessing disparity when protected class is unobserved. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3287560.3287594

Chen, John, Berlot-Attwell, I., Wang, X., Hossain, S., & Rudzicz, F. (2020). *Exploring Text Specific and Blackbox Fairness Algorithms in Multimodal Clinical NLP*. https://doi.org/10.18653/v1/2020.clinicalnlp-1.33

Chiappa, S. (2019). Path-specific counterfactual fairness. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*. https://doi.org/10.1609/aaai.v33i01.33017801

Choi, Y., Dang, M., & van den Broeck, G. (2021a). *Group Fairness by Probabilistic Modeling with Latent Fair Decisions*. Retrieved from www.aaai.org

Choi, Y., Dang, M., & van den Broeck, G. (2021b). *Group Fairness by Probabilistic Modeling with Latent Fair Decisions*. Retrieved from www.aaai.org

Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, *5*(2). https://doi.org/10.1089/big.2016.0047

Chu, X., Zhang, B., & Xu, R. (2022). *FairNAS: Rethinking Evaluation Fairness of Weight Sharing Neural Architecture Search*. https://doi.org/10.1109/iccv48922.2021.01202

Claure, H., Chen, Y., Modi, J., Jung, M., & Nikolaidis, S. (2020). Multi-armed bandits with fairness constraints for distributing resources to human teammates. *ACM/IEEE International Conference on Human-Robot Interaction*. https://doi.org/10.1145/3319502.3374806

COMPAS (software) - Wikipedia. (n.d.). Retrieved May 14, 2022, from https://en.wikipedia.org/wiki/COMPAS_(software)

Corcoran, E., Denman, S., & Hamilton, G. (2021). Evaluating new technology for biodiversity monitoring: Are drone surveys biased? *Ecology and Evolution*, *11*(11). https://doi.org/10.1002/ece3.7518

Correa, J., Cristi, A., Duetting, P., & Norouzi-Fard, A. (2021). Fairness and Bias in Online Selection. *Proceedings of the 38th International Conference on Machine Learning*, *139*.

Craven, M. W., & Shavlik, J. W. (1996). Extracting tree-structured representations of trained neural networks. *Advances in Neural Information Processing Systems*, *8*.

del Barrio, E., Gamboa, F., Gordaliza, P., & Loubes, J. M. (2019). Obtaining fairness using optimal transport theory. *36th International Conference on Machine Learning, ICML 2019, 2019-June*.

Dong, Y., Kang, J., Tong, H., & Li, J. (2021). Individual Fairness for Graph Neural Networks: A Ranking based Approach. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/3447548.3467266

Du, M., Mukherjee, S., Wang, G., Tang, R., Awadallah, A. H., & Hu, X. (2021). *Fairness via Representation Neutralization*.

Du, X., Pei, Y., Duivesteijn, W., & Pechenizkiy, M. (2020). Fairness in network representation by latent structural heterogeneity in observational data. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*. https://doi.org/10.1609/aaai.v34i04.5792

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *ITCS 2012 - Innovations in Theoretical Computer Science Conference*. https://doi.org/10.1145/2090236.2090255

Dybå, T., Kitchenham, B. A., & Jorgensen, M. (2005). Evidence-based software engineering for practitioners. *IEEE Software*, *22*(1). https://doi.org/10.1109/MS.2005.6

Franco, D., Oneto, L., Navarin, N., & Anguita, D. (2021). Toward learning trustworthily from data combining privacy, fairness, and explainability: An application to face recognition. *Entropy*, *23*(8). https://doi.org/10.3390/e23081047

Galhotra, S., Shanmugam, K., Sattigeri, P., & Varshney, K. R. (2021). Interventional fairness with indirect knowledge of unobserved protected attributes. *Entropy*, *23*(12). https://doi.org/10.3390/e23121571

Goel, N., Yaghini, M., & Faltings, B. (2018). Non-discriminatory machine learning through convex fairness criteria. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*. https://doi.org/10.1145/3278721.3278722

Google apologises for Photos app's racist blunder - BBC News. (n.d.). Retrieved May 12, 2022, from https://www.bbc.com/news/technology-33347866

Grari, V., Hajouji, O. el, Lamprier, S., & Detyniecki, M. (2021). Learning Unbiased Representations via Rényi Minimization. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12976 LNAI*. https://doi.org/10.1007/978-3-030-86520-7_46

Grgic-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2016). The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making. In *NIPS symposium on machine learning and the law*.

Griggs v. Duke Power Co. :: 401 U.S. 424 (1971) :: Justia US Supreme Court Center. (n.d.). Retrieved May 15, 2022, from https://supreme.justia.com/cases/federal/us/401/424/

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*.

Hiring Practices That Have a Negative Effect on Certain Applicants | U.S. Equal Employment Opportunity Commission. (n.d.). Retrieved May 15, 2022, from https://www.eeoc.gov/employers/small-business/hiring-practices-have-negative-effect-certain-applicants

Holmes, E. (2005). Anti-Discrimination Rights Without Equality. *Modern Law Review*, *68*(2). https://doi.org/10.1111/j.1468-2230.2005.00534.x

Hu, H., Liu, Y., Wang, Z., & Lan, C. (2019). A distributed fair machine learning framework with private demographic data protection. *Proceedings - IEEE International Conference on Data Mining, ICDM*, *2019-November*. https://doi.org/10.1109/ICDM.2019.00131

Hu, T., Iosifidis, V., Liao, W., Zhang, H., Yang, M. Y., Ntoutsi, E., & Rosenhahn, B. (2020). FairNN - Conjoint Learning of Fair Representations for Fair Decisions. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12323 LNAI*. https://doi.org/10.1007/978-3-030-61527-7_38

Hwang, S., Park, S., Kim, D., & Byun, H. (2020). *FairFaceGAN: Fairness-aware Facial Image-to-Image Translation Mirae Do*.

Hwang, S., Park, S., Lee, P., Jeon, S., Kim, D., & Byun, H. (2021). Exploiting Transferable Knowledge for Fairness-Aware Image Classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12625 LNCS*. https://doi.org/10.1007/978-3-030-69538-5_2

Iosifidis, V., Fetahu, B., & Ntoutsi, E. (2019). FAE: A Fairness-Aware Ensemble Framework. *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*. https://doi.org/10.1109/BigData47090.2019.9006487

Iosifidis, V., & Ntoutsi, E. (2020). - Online Fairness-Aware Learning Under Class Imbalance. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12323 LNAI*, 159–174. https://doi.org/10.1007/978-3-030-61527-7_11

Iosifidis, V., Tran, T. N. H., & Ntoutsi, E. (2019). Fairness-Enhancing Interventions in Stream Classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11706 LNCS*. https://doi.org/10.1007/978-3-030-27615-7_20

ISO - ISO 26000 — Social responsibility. (n.d.). Retrieved May 12, 2022, from https://www.iso.org/iso-26000-social-responsibility.html

Kassoff, A., Kassoff, J., Buehler, J., Eglow, M., Kaufman, F., Mehu, M., … Crouse, V. D. (2001). A randomized, placebo-controlled, clinical trial of high-dose supplementation with vitamins C and E, beta carotene, and zinc for age-related macular degeneration and vision loss: AREDS report no. 8. *Archives of Ophthalmology*, *119*(10). https://doi.org/10.1001/archopht.119.10.1417

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *35th International Conference on Machine Learning, ICML 2018*, *6*.

Kearns, M., Roth, A., Neel, S., & Wu, Z. S. (2019). An empirical study of rich subgroup fairness for machine learning. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3287560.3287592

Kim, M. P., Ghorbani, A., & Zou, J. (2019). Multiaccuracy: Black-box post-processing for fairness in classification. *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1145/3306618.3314287

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., … Reblitz-Richardson, O. (2020). *Captum: A unified and generic model interpretability library for PyTorch*.

Krasanakis, E., Spyromitros-Xioufis, E., Papadopoulos, S., & Kompatsiaris, Y. (2018). Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018*. https://doi.org/10.1145/3178876.3186133

Krause, J., Perer, A., & Ng, K. (2016). Interacting with predictions: Visual inspection of black-box machine learning models. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/2858036.2858529

Kusner, M., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in Neural Information Processing Systems*, *2017-December*.

Liu, S., & Nunes Vicente, L. (2020). *Accuracy and Fairness Trade-offs in Machine Learning: A Stochastic Multi-Objective Approach*.

Liu, W., Liu, F., Tang, R., Liao, B., Chen, G., & Heng, P. A. (2020). Balancing Between Accuracy and Fairness for Interactive Recommendation with Reinforcement Learning. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12084 LNAI*. https://doi.org/10.1007/978-3-030-47426-3_13

Lohia, P. K., Natesan Ramamurthy, K., Bhide, M., Saha, D., Varshney, K. R., & Puri, R. (2019). Bias Mitigation Post-processing for Individual and Group Fairness. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2019-May*. https://doi.org/10.1109/ICASSP.2019.8682620

Louizos, C., Swersky, K., Li, Y., Welling, M., & Zemel, R. (2016). The variational fair autoencoder. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 2017-December*.

Lyu, L., He, X., & Li, Y. (2020). Differentially private representation for NLP: Formal guarantee and an empirical study on privacy and fairness. *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*. https://doi.org/10.18653/v1/2020.findings-emnlp.213

Lyu, L., Xu, X., Wang, Q., & Yu, H. (2020). Collaborative Fairness in Federated Learning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*: *Vol. 12500 LNCS*. https://doi.org/10.1007/978-3-030-63076-8_14

Ma, J., Deng, J., & Mei, Q. (2021). *Subgroup Generalization and Fairness of Graph Neural Networks*. Retrieved from https://github.com/TheaperDeng/GNN-Generalization-Fairness.

Machine Bias — ProPublica. (n.d.). Retrieved May 12, 2022, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Madras, D., Creager, E., Pitassi, T., & Zemel, R. (2019). Fairness through causal awareness: Learning causal latent-variable models for biased data. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3287560.3287564

Madras, D., Pitassi, T., & Zemel, R. (2018). Predict responsibly: Improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems*, *2018-December*.

Mandal, D., Deng, S., Jana, S., Wing, J. M., & Hsu, D. (2020). Ensuring fairness beyond the training data. *Advances in Neural Information Processing Systems*, *2020-December*.

Marion, D., Funda, C. E., Peter, B., Francesca, C., Peter, H., Stylianos, K., … Jacques, D. F. (2017). *What makes a fair society? Insights and evidence*.

Mccarthy, J. (2007). *WHAT IS ARTIFICIAL INTELLIGENCE?* Retrieved from http://www-formal.stanford.edu/jmc/

Mireshghallah, F., & Berg-Kirkpatrick, T. (2021). *Style Pooling: Automatic Text Style Obfuscation for Improved Classification Fairness*. https://doi.org/10.18653/v1/2021.emnlp-main.152

Nandy, P., Diciccio, C., Venugopalan, D., Logan, H., Basu, K., & Karoui, N. el. (2020). *Achieving Fairness via Post-Processing in Web-Scale Recommender Systems*.

Noriega-Campero, A., Garcia-Bulle, B., Bakker, M. A., & Pentland, A. S. (2019). Active fairness in algorithmic decision making. *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1145/3306618.3314277

Noroozi, V., Bahaadini, S., Sheikhi, S., Mojab, N., & Yu, P. S. (2019). Leveraging semi-supervised learning for fairness using neural networks. *Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019*. https://doi.org/10.1109/ICMLA.2019.00017

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

Open data and data bias | data.europa.eu. (n.d.). Retrieved May 14, 2022, from https://data.europa.eu/en/news/open-data-and-data-bias

Park, S., Hwang, S., Kim, D., & Byun, H. (2021). Learning Disentangled Representation for Fair Facial Attribute Classification via Fairness-aware Information Alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(3).

Paul, W., Hadzic, A., Joshi, N., & Burlina, P. (2020). RENATA: REpreseNtation And Training Alteration for Bias Mitigation. *Preprint*, *1*(1).

Pedreshi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/1401890.1401959

Peng, K., Chakraborty, J., & Menzies, T. (2021). *IEEE TRANSACTIONS ON SOFTWARE ENGINEERING 1 xFAIR: Better Fairness via Model-based Rebalancing of Protected Attributes*. Retrieved from https://github.com/anonymous12138/biasmitigation.

Petersen, F., Mukherjee, D., Sun, Y., & Yurochkin, M. (n.d.). *Post-processing for Individual Fairness*.

Preot'iuc-Pietro, D., Lampos, V., & Aletras, N. (2015). An analysis of the user occupational class through Twitter content. *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian*

*Federation of Natural Language Processing, Proceedings of the Conference*, *1*. https://doi.org/10.3115/v1/p15-1169

Pruksachatkun, Y., Krishna, S., Dhamala, J., Gupta, R., & Chang, K. W. (2021). Does Robustness Improve Fairness? Approaching Fairness with Word Substitution Robustness Methods for Text Classification. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. https://doi.org/10.18653/v1/2021.findings-acl.294

Puyol-Antón, E., Ruijsink, B., Piechnik, S. K., Neubauer, S., Petersen, S. E., Razavi, R., & King, A. P. (2021). Fairness in Cardiac MR Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12903 LNCS*. https://doi.org/10.1007/978-3-030-87199-4_39

Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures | U.S. Equal Employment Opportunity Commission. (n.d.). Retrieved May 15, 2022, from https://www.eeoc.gov/laws/guidance/questions-and-answers-clarify-and-provide-common-interpretation-uniform-guidelines

Ranzato, F., Urban, C., & Zanella, M. (2021). Fairness-Aware Training of Decision Trees by Abstract Interpretation. *International Conference on Information and Knowledge Management, Proceedings*. https://doi.org/10.1145/3459637.3482342

Ravichandran, S., Khurana, D., Labs, A., Express Bangalore, A., Bharath Venkatesh, K., Unny Edakunni, N., & Venkatesh, B. (2020). *FairXGBoost: Fairness-aware Classification in XGBoost*. https://doi.org/10.1145/1122445.1122456

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). *Model-Agnostic Interpretability of Machine Learning*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD*

*International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*. https://doi.org/10.1145/2939672.2939778

Roh, Y., Lee, K., Whang, S. E., & Suh, C. (2021). *FAIRBATCH: BATCH SELECTION FOR MODEL FAIRNESS*.

Rudin, C. (2013). Predictive policing: Using machine learning to detect patterns of crime. *Ecml Pkdd*.

Sadeghi, B., & Boddeti, V. N. (2020). Imparting fairness to pre-trained biased representations. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, *2020-June*. https://doi.org/10.1109/CVPRW50498.2020.00016

Sarhan, M. H., Navab, N., Eslami, A., & Albarqouni, S. (2020). Fairness by Learning Orthogonal Disentangled Representations. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12374 LNCS*. https://doi.org/10.1007/978-3-030-58526-6_44

Sattigeri, P., Hoffman, S. C., Chenthamarakshan, V., & Varshney, K. R. (2019). Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, *63*(4–5). https://doi.org/10.1147/JRD.2019.2945519

Sharma, S., Henderson, J., & Ghosh, J. (2020). CERTIFAI: A common framework to provide explanations and analyse the fairness and robustness of black-box models. *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1145/3375627.3375812

Shekhar, S., Shah, N., & Akoglu, L. (2021). FairOD: Fairness-aware Outlier Detection. *AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1145/3461702.3462517

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *34th International Conference on Machine Learning, ICML 2017*, *7*.

Singh, A., & Joachims, T. (2019). Policy learning for fairness in ranking. *Advances in Neural Information Processing Systems*, *32*.

Singh, H., Singh, R., Mhasawade, V., & Chunara, R. (2021). Fairness violations and mitigation under covariate shift. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3442188.3445865

Slack, D., Friedler, S. A., & Givental, E. (2020). Fairness warnings and Fair-maml: Learning fairly with minimal data. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3351095.3372839

Slob, N., Catal, C., & Kassahun, A. (2021). Application of machine learning to improve dairy farm management: A systematic literature review. *Preventive Veterinary Medicine*, Vol. 187. https://doi.org/10.1016/j.prevetmed.2020.105237

Spape, M., Davis, K., Kangassalo, L., Ravaja, N., Sovijarvi-Spape, Z., & Ruotsalo, T. (2021). Brain-computer interface for generating personally attractive images. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2021.3059043

Spinelli, I., Scardapane, S., Hussain, A., & Uncini, A. (2021). FairDrop: Biased Edge Dropout for Enhancing Fairness in Graph Representation Learning. *IEEE Transactions on Artificial Intelligence*. https://doi.org/10.1109/tai.2021.3133818

Srivastava, M., Heidari, H., & Krause, A. (2019). Mathematical notions vs. Human perception of fairness: A descriptive approach to fairness for machine learning. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/3292500.3330664

Stoyanovich, J., Yang, K., & Jagadish, H. v. (2018). Online set selection with fairness and diversity constraints. *Advances in Database Technology - EDBT*, *2018-March*. https://doi.org/10.5441/002/edbt.2018.22

Štrumbelj, E., & Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, *11*.

Subramanian, S., Rahimi, A., Baldwin, T., Cohn, T., & Frermann, L. (2021). *Fairness-aware Class Imbalanced Learning*. https://doi.org/10.18653/v1/2021.emnlp-main.155

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *34th International Conference on Machine Learning, ICML 2017, 7*.

Svoboda, E. (2020). Artificial intelligence is improving the detection of lung cancer. *Nature*, Vol. 587. https://doi.org/10.1038/d41586-020-03157-9

Theil Index. (n.d.). Retrieved May 13, 2022, from https://www.census.gov/topics/income-poverty/income-inequality/about/metrics/theil-index.html

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., … Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, *596*(7873). https://doi.org/10.1038/s41586-021-03828-1

Valentim, I., Lourenco, N., & Antunes, N. (2019). The Impact of Data Preparation on the Fairness of Software Systems. *Proceedings - International Symposium on Software Reliability Engineering, ISSRE, 2019-October*. https://doi.org/10.1109/ISSRE.2019.00046

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., … Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, *575*(7782). https://doi.org/10.1038/s41586-019-1724-z

Wagner, B., & d'Avila Garcez, A. (2021). Neural-symbolic integration for fairness in AI. *CEUR Workshop Proceedings*, *2846*.

Wang, T., & Saar-Tsechansky, M. (2020). *Augmented Fairness: An Interpretable Model Augmenting Decision-Makers' Fairness*.

Wang, Y., Wang, X., Beutel, A., Prost, F., Chen, J., & Chi, E. H. (2021). Understanding and Improving Fairness-Accuracy Trade-offs in Multi-Task Learning. *Proceedings of the ACM SIGKDD International Conference on*

*Knowledge Discovery and Data Mining*. https://doi.org/10.1145/3447548.3467326

Wang, Zeyu, Qinami, K., Karakozis, I. C., Genova, K., Nair, P., Hata, K., & Russakovsky, O. (2020). Towards fairness in visual recognition: Effective strategies for bias mitigation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR42600.2020.00894

Wang, Zhao, Shu, K., & Culotta, A. (2021). *Enhancing Model Robustness and Fairness with Causality: A Regularization Approach*. https://doi.org/10.18653/v1/2021.cinlp-1.3

Xiao, L., Nouri, S., Chapman, M., Fix, A., Lanman, D., & Kaplanyan, A. (2020). Neural supersampling for real-time rendering. *ACM Transactions on Graphics*, *39*(4). https://doi.org/10.1145/3386569.3392376

Xing, X., Liu, H., Chen, C., & Li, J. (2021). Fairness-Aware Unsupervised Feature Selection. *International Conference on Information and Knowledge Management, Proceedings*. https://doi.org/10.1145/3459637.3482106

Xu, D., Wu, Y., Yuan, S., Zhang, L., & Wu, X. (2019). Achieving causal fairness through generative adversarial networks. *IJCAI International Joint Conference on Artificial Intelligence, 2019-August*. https://doi.org/10.24963/ijcai.2019/201

Xu, D., Yuan, S., & Wu, X. (2019). Achieving differential privacy and fairness in logistic regression. *The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019*. https://doi.org/10.1145/3308560.3317584

Xu, D., Yuan, S., Zhang, L., & Wu, X. (2019). FairGAN: Fairness-aware Generative Adversarial Networks. *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*. https://doi.org/10.1109/BigData.2018.8622525

Xu, T., White, J., Kalkan, S., & Gunes, H. (2020). Investigating Bias and Fairness in Facial Expression Recognition. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12540 LNCS*. https://doi.org/10.1007/978-3-030-65414-6_35

Xu, Xingkun, Huang, Y., Shen, P., Li, S., Li, J., Huang, F., … Cui, Z. (2021). Consistent instance false positive improves fairness in face recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR46437.2021.00064

Xu, Xinyi, Lyu, L., Ma, X., Miao, C., Foo, C. S., & Low, B. K. H. (2021). Gradient Driven Rewards to Guarantee Fairness in Collaborative Machine Learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, & J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems* (Vol. 34, pp. 16104–16117). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2021/file/8682cc30db9c025ecd3fee433f8ab54c-Paper.pdf

Yang, F., Cisse, M., & Koyejo, S. (2020). Fairness with overlapping groups. *Advances in Neural Information Processing Systems*, *2020-December*.

Yoon, T., Lee, J., & Lee, W. (2020). Joint Transfer of Model Knowledge and Fairness over Domains Using Wasserstein Distance. *IEEE Access*, *8*. https://doi.org/10.1109/ACCESS.2020.3005987

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). *Fairness Beyond Disparate Treatment & Disparate Impact*. https://doi.org/10.1145/3038912.3052660

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. *30th International Conference on Machine Learning, ICML 2013*, (PART 2).

Zhang, T., Zhu, T., Li, J., Han, M., Zhou, W., & Yu, P. S. (2022). Fairness in Semi-Supervised Learning: Unlabeled Data Help to Reduce Discrimination. *IEEE Transactions on Knowledge and Data Engineering*, *34*(4). https://doi.org/10.1109/TKDE.2020.3002567

Zhang, W., & Bifet, A. (2020). FEAT: A Fairness-Enhancing and Concept-Adapting Decision Tree Classifier. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12323 LNAI*. https://doi.org/10.1007/978-3-030-61527-7_12

Zhang, W., Bifet, A., Zhang, X., Weiss, J. C., & Nejdl, W. (2021). FARF: A Fair and Adaptive Random Forests Classifier. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12713 LNAI*. https://doi.org/10.1007/978-3-030-75765-6_20

Zhang, W., & Ntoutsi, E. (2019). FaHT: An adaptive fairness-aware decision tree classifier. *IJCAI International Joint Conference on Artificial Intelligence*, *2019-August*. https://doi.org/10.24963/ijcai.2019/205

Zhang, Yifu, Wang, C., Wang, X., Zeng, W., & Liu, W. (2021). FairMOT: On the Fairness of Detection and Re-identification in Multiple Object Tracking. *International Journal of Computer Vision*, *129*(11). https://doi.org/10.1007/s11263-021-01513-4

Zhang, Yue, & Ramesh, A. (2020). Learning fairness-aware relational structures. *Frontiers in Artificial Intelligence and Applications*, *325*. https://doi.org/10.3233/FAIA200389

Zhao, B., Xiao, X., Gan, G., Zhang, B., & Xia, S. (2020). Maintaining discrimination and fairness in class incremental learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR42600.2020.01322

Zhao, C., Chen, F., & Thuraisingham, B. (2021). Fairness-Aware Online Meta-learning. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/3447548.3467389

Zhao, F., Huang, Y., Maradapu Vera Venkata Sai, A., & Wu, Y. (2020, May). *A Cluster-based Solution to Achieve Fairness in Federated Learning*. 875–882. https://doi.org/10.1109/ISPA-BDCloud-SocialCom-SustainCom51426.2020.00135

Zhu, Z., Hu, X., & Caverlee, J. (2018). Fairness-aware tensor-based recommendation. *International Conference on Information and Knowledge Management, Proceedings*. https://doi.org/10.1145/3269206.3271795