Probability-based Algorithmic Framework Incorporating Human-centric Design and Implicit Bias to Counter Discriminatory Predictions by AI/ML Systems

Tanish Kelkar
A.D Joshi Junior College
Maharashtra, India
kelkartanish@gmail.com

Ashug Gurijala
Oakridge International School, Newton Campus
Telangana, India
ashuggurijala2@gmail.com

Shizhe He
Otto-von-Taube-Gymnasium
Gauting, Germany
shizhe.he@tum.de

Kian Andrew Busico
Philippine Science High School - Southern Mindanao Campus
Davao City, Philippines
kabusico@gmail.com

Avni Gupta
Wilton High School
Connecticut, USA
avnig2005@gmail.com

Abstract

Machine learning models have been employed in many organizations for impactful decision-making processes. However, bias prevalent in an organization's model could create meaningful issues, ranging from incorrect hiring decisions to a negative impact on an organization's image. Organizations often lack the knowledge or time to create a model in line with society's standards, leading to the utilization of faulty systems. Here, we present a framework to help combat this issue, allowing organizations to prevent bias in their models. Our framework starts with a general guidebook to clarify relevant terminology and provide best practices for developers to measure and mitigate bias. We then incorporate an automated algorithm based on statistical parity and disparate impact to de-bias raw data. We apply SMOTE-Tomek to resample imbalanced datasets and a Reject Option Classification algorithm to reduce biased predictions. We infer adding more data related to minority and unprivileged classes to datasets will help create more equitable and representative datasets leading to fair AI systems for all. We propose a novel decentralized database using web-scraping and homomorphic encryption as a reliable source of real-world data. Our assumed hypotheses were confirmed by data derived from extensive testing.

Keywords

Artificial Intelligence, Machine Learning, Implicit Bias, Homomorphic encryption, Web Scraping

1. Introduction

From content recommendations to predicting recidivism, Artificial Intelligence influences almost all facets of life. Over 67% of leading businesses rely heavily on Artificial Intelligence systems. Investment in AI systems has increased sixfold since 2020, indicating that AI systems will be even more widely adopted. AI Systems are projected to increase at a rate of 247% over the next 5 years, including in high-stake fields like healthcare, justice, and recruiting. This implies that several life-altering decisions like patient diagnosis and treatment, hail terms, hiring decisions, etc. will be influenced by AI algorithms.

We have noted many past incidents of bias in AI systems, making apparent that various forms of bias are present in AI systems across numerous sectors. Furthermore, over 83% of AI Algorithms in use today are biased against people of a particular group. A data scientist survey by CrowdFlower revealed that 63% of data scientists surveyed listed biased AI as the top concern in their field. A mere Google search for 'healthy skin' shows images of predominantly light-skinned women, whereas negative search terms linked to drug addiction mostly reveal photos of dark-skinned individuals. Several reports have shown that the COMPAS algorithm favors white men over black. Considering the growing prevalence of AI as a tool, bias in AI has a high potential to exacerbate already-existing societal disparities.

Biased training data is the principal cause for Biased AI Systems. Even if the data is sourced accurately, it can contain historical human bias which can lead the classifier algorithm to form incorrect or biased assumptions leading to discriminatory predictions. Our research also points to the fact that imbalanced datasets can lead to bias against minority groups. The "Faces in the Wild" dataset, which consists of 80% white and 70% male faces, led to significantly biased predictions by systems reliant on it.

One of the major hurdles addressed by the research herein is defining bias Bias is a fairly complicated and context-based multi-faceted concept. Narayanan described at least 21 mathematical definitions of fairness (Narayanan, 2018). Different definitions produce entirely different outcomes. For example, ProPublica and Northpointe had a public debate on an important social justice issue (recidivism prediction) that was fundamentally about what is the right fairness metric (Larson et al., 2016; Dieterich et al., 2016; Larson & Angwin, 2016). Furthermore, researchers have shown that it is impossible to satisfy all definitions of fairness at the same time (Kleinberg et al., 2017). To solve the ambiguity, the paper presents a clear definition of fairness and types of Bias in Raw Data. A flowchart graphic is designed to outline the best fairness metric for the user.

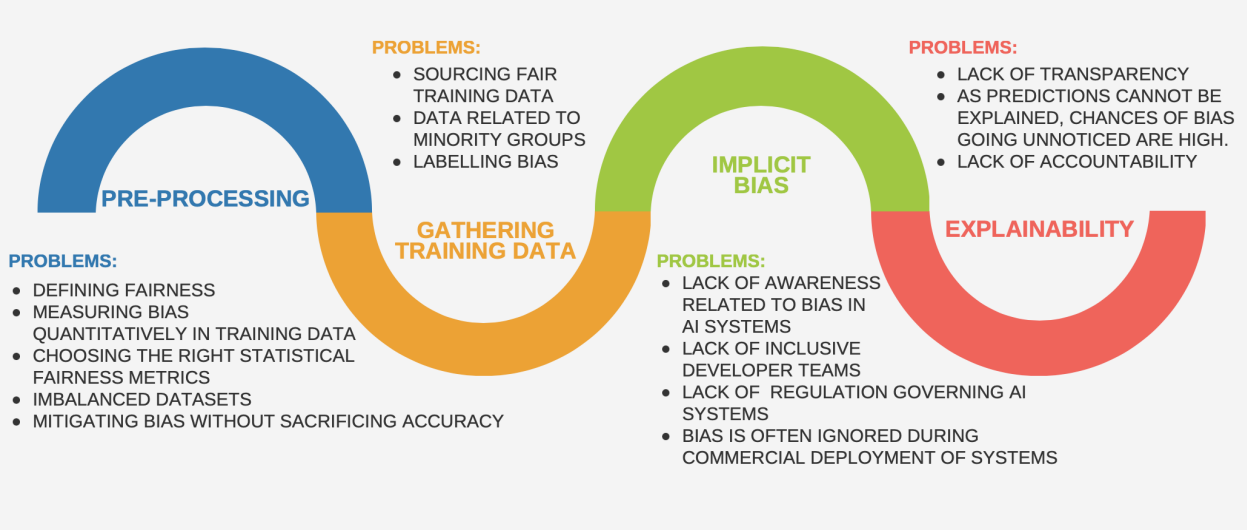

Figure 1. Problems identified in four stages of a Machine Learning pipeline

To solve the problem of discriminatory predictions, we have chosen four stages of the machine learning pipeline: Pre-Processing, Gathering Training Data, Implicit Bias, and Explainability. Figure 1 illustrates the list of problems that this research has identified and intends to solve in each stage. The Development of an AI System is a complex process with multiple steps, all of which can introduce bias into the final product. We have chosen to isolate problems in the

steps of developing an AI system, targeting both the technical and human aspects of development. Our solutions address the technical, social, and economic aspects of each highlighted problem.

1.1 Objectives

- To Define fairness and types of bias in raw data
- Propose a simple, human-centric software framework that eliminates bias in pre-processing.
- Establish that imbalanced datasets can lead to discriminatory predictions, which may pose harm
- Use the SMOTE-TOMEK algorithm in tandem with the software to create an all-inclusive mitigation tool for developers.
- Establish that Implicit Bias is an indirect contributor to Biased AI Systems and design solutions to counter Implicit Bias.

2 Application

This chapter elaborates on our bias measuring and mitigating platform (https://github.com/h3seas0n/aifairness_ja). This core part of our solution will significantly reduce bias in training data and therefore, will help create fair AI systems.

The Development of an AI System is a complex process with multiple steps, all of which can introduce bias into the final product. Most of the bias analysis and mitigation techniques involve the creation of separate teams and the deployment of valuable resources. Therefore, Bias Analysis and Mitigation is usually ignored and is also out of hand for small-scale companies and developers without huge institutional and financial resources. We have isolated problems in each step of developing a deployable machine learning model. By developing this application, we strive to address specific issues identified in every part of the machine learning pipeline in a user-friendly, seamless process. Consisting of several overview panels and codeless algorithm input pages, our bias measuring and mitigation platform provides an all-inclusive, scalable, and pragmatic solution for both researchers/research groups and end-users such as local governments or companies with and without previous experience in machine learning. Additionally, our solution makes it incredibly easy for AI developers to test and deploy numerous fairness metrics as well as the pre-processing mitigation algorithms on one single platform with ease. This way, we provide a highly efficient and effective solution for our users to measure and mitigate bias in their datasets:
To create this easy-to-use bias mitigation platform, our team at AI404 developed an electron-like application with a frontend based on HTML, CSS, JavaScript, and a Python-based backend. We utilized the python library “eel” to connect these two parts and implement the bias measuring and mitigation algorithms while maintaining a clean, modern graphical user interface. Using the Bootstrap 4 framework, we were able to build a modern, responsive frontend interface with several input components including modal popups and accordion dropdown selections. Connected with the bootstrap 4 frontend, the python-based backend employs several components of the AI Fairness 360 library to measure and mitigate data and algorithmic bias based on the user’s input passed through eel. Through bootstrap and eel, the user can upload their own custom dataset and define all parameters needed for our vast range of bias metrics and mitigation algorithms.

2.1 Pre-Processing

Pre-Processing bias mitigation is targeted at bias prevalent in the training data before model training. Our comprehensive bias testing and mitigation platform will allow users to measure bias in training data using over 40 state-of-the-art fairness metrics and mitigate bias through a reweighting algorithm. The solution will also mitigate imbalanced datasets by resampling the majority and minority classes using advanced algorithms like SMOTE+Tomek and Synthetic Data generators (GAN's).

2.1.1 Bias in Training Data

Raw data bias can be perceived when datasets contain predefined prejudice towards certain groups based on sensitive categories (including age, disability, marital status, maternity, race, religious belief, sexual orientation, and sex). A biased dataset does not accurately represent a model's use case, resulting in skewed outcomes, low accuracy levels, and analytical errors. Biased data also includes content produced by humans which may contain bias against groups of people (see Fig. 2).

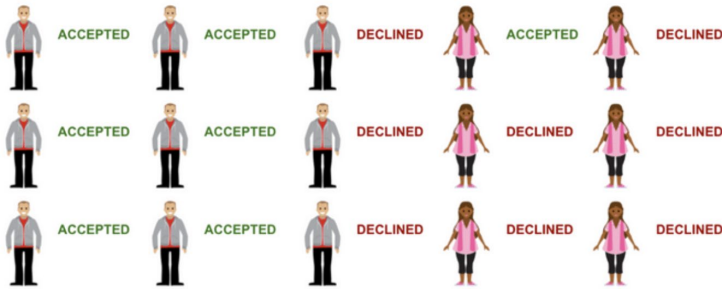

Figure 2: A dataset with uneven representation

If a model is trained on the training data illustrated in Figure 2, It may lead to biased predictions against women. Most biases found in machine learning datasets can be assigned to one of four categories: Sampling bias, Association Bias, Negative Feedback, and Observer/Labelling Bias. Sample bias occurs when a dataset does not reflect the realities of the environment in which a model will run, whereas cultural bias occurs when the data for a machine learning model reinforces and/or multiplies a cultural bias. The machine learning model itself can also generate bias itself by influencing the generation of data that is used to train it, known as negative feedback - A compelling example of this is the controversial Microsoft Tay ChatBot. Apart from this, the developer of the machine learning algorithm is able to introduce observer bias, also known as confirmation bias, into their models by developing with an outcome mind, and thus fail to see deviations from that outcome.

Another form of bias in datasets is inadequate representation (when the majority of the dataset (or even a subset) is heavily skewed towards one conclusion or the other). Image labeling also plays a role, as systems trained on images labeled with derogatory terms will implement those terms. Additionally, data reflecting historical trends is more often than not biased, since the status quo is already biased. If historical data is used, it only exacerbates the effects of bias in those areas.

However, it is not efficient nor effective to remove certain attributes to mitigate bias due to the potential dependency of the algorithm on certain sensitive variables. It has been proven in several papers that fairness through unawareness leads only to lower accuracy and higher bias. There are several steps and methods to reduce data bias significantly. Apart from social exercises and reliable data-collection services, several algorithms can be utilized to measure and mitigate bias in raw data before model training.

2.1.2 Fairness Metrics

A dataset can be said to be fair if the results of an AI algorithm trained on it are independent of certain parameters that should not affect the outcome - These variables are referred to as sensitive attributes (gender, ethnicity, sexual orientation, etc.). After thorough research, we concluded that the best and the most accurate way to quantitatively evaluate bias in training data was to use Fairness metrics. There are several existing research studies, including those by IBM, Facebook, and Google, that have proven with evidence that fairness metrics measure bias in data accurately and reliably.

Several criteria can be used to assess whether the trained algorithm is fair based on the application of the AI algorithm to be trained. For example, statistical parity, being one of the most well-known criteria, is satisfied when the sensitive attributes of the input are statistically independent of the model output. This is to say, two individuals with the same characteristics apart from the protected attribute (sensitive characteristic) have the same possibility of being characterized as the same output. However, we have to keep in mind that there are many other criteria with certain advantages and flaws that one can utilize to determine fairness in different use cases - Choosing the right fairness criteria is extremely subjective and dependent on the stakeholders. As proved by Kleinberg et.al, satisfying all metrics is impossible. However, we can choose to optimize the ones most applicable to the use case and the stakeholders of the AI System (see Figure 3):

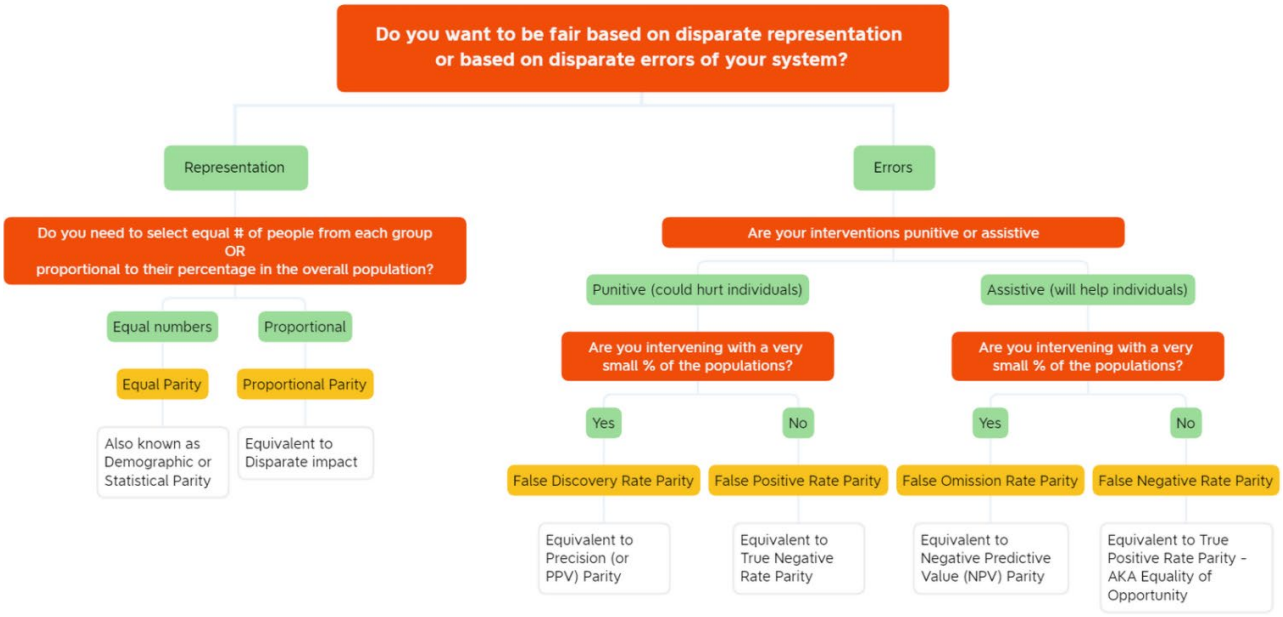

Figure 3: A flowchart to help our users understand and target the most applicable fairness metrics.

The AI404's Comprehensive Bias Measuring and Mitigation Platform utilizes a variety of fairness metrics to evaluate Bias in Training Data. Most of the Fairness Metrics in our solution, except disparate impact and statistical parity, are calculated using a confusion matrix (See Figure 4). For instance, the Average Odds Difference metric is calculated as the average difference of the false-positive rate (false-positives/negatives) and the true-positive rate (true positives/positives) between unprivileged and privileged groups. If the value of this metric is above 0.1 or below -0.1, it proves that the dataset is biased:

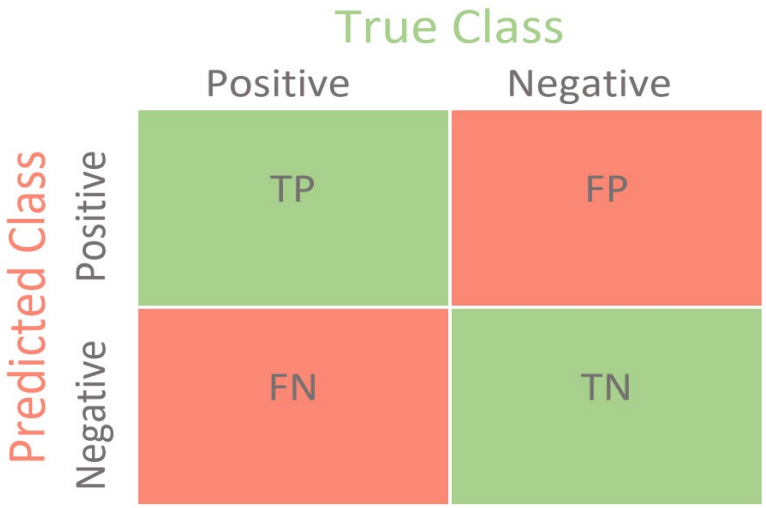

Figure 4: Confusion Matrix used to derive most of the fairness metrics

$$Average\ Odds\ Difference\ metric = \frac{1}{2}\left[(FPR_{D=unprivileged} - FPR_{D=privileged}) + (TPR_{D=unprivileged} - TPR_{D=privileged})\right]$$

Several mitigation algorithms target optimizing certain metric values. An example of this is reweighing, a simple methodology for mitigating bias. Instead of changing certain attribute values, the algorithm generates weights (form: vector) for each output instance based on the protected attribute, the favorable label, and the protected attribute which is further passed to the model: By applying the output weights to the data, the disparate impact metric, which can be used to measure data bias, becomes 1, indicating discrimination-free data. However, algorithms for bias mitigation

are not merely limited to simple reweighing. Further algorithms such as Optimum pre-processing and Learning fair representations (LFR) enable usage in many more use-cases, increasing the possible impact on AI systems.

2.2 Resampling

The bias in a training dataset influences the performance of the machine learning algorithms heavily, as algorithms amplify patterns and often ignore minority classes. Generally, however, predictions for minority classes may be classified as more important, but their underrepresentation in datasets does not allow algorithms to create accurate predictions. The bias in the training dataset occurs due to imbalanced data where there is a severe skew in the class distribution, sometimes in the ratio of 1:100 for minority to majority classes. Several techniques may be used to address this imbalanced data. The techniques that will be researched use either oversampling (add more samples to minority class) or undersampling (remove samples for majority class).

2.2.1 Random Over/Undersampling

In Random Oversampling, the examples of the minority class are randomly selected with replacement and added to the training dataset. As the name indicates, the examples are selected randomly without using any heuristics. This makes the process simple and fast, and therefore widely used for large and complex datasets. However, Random Oversampling increases the likelihood of occurrence of overfitting, since it makes exact copies of the minority class examples. The effect can lead to performance on the training dataset, but worse performance on the test dataset. Figure 5 illustrates how an imbalance dataset can be resampled using random oversampling.

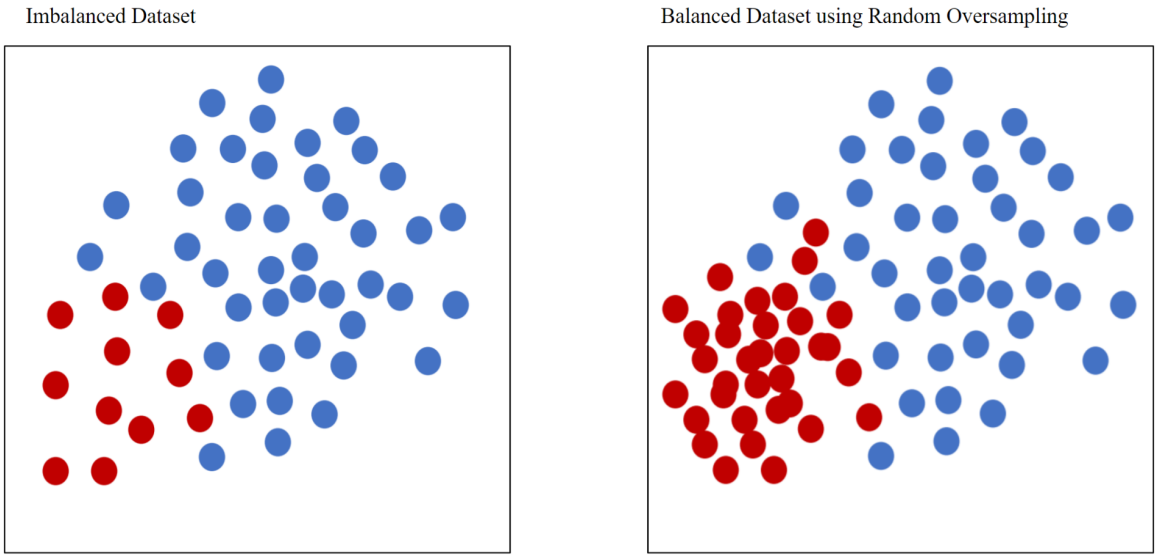

Figure 5. Balancing data using random oversampling

Similarly, in random undersampling, the examples of the majority class are randomly removed from the training dataset. This process is also simple and fast, but a major drawback of this method is that potentially useful data could be discarded, when it may be important for the induction process. Figure 6 illustrates a balanced dataset, after resampling the imbalanced dataset in Figure 5 using random undersampling.

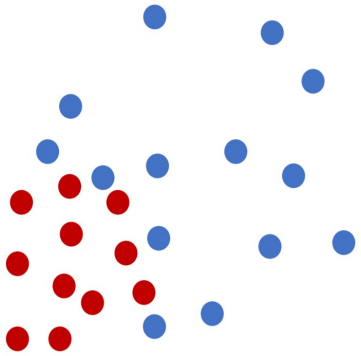

Figure 6. A balanced dataset using random oversampling

2.2.2 SMOTE+Tomek Links

Synthetic Minority Oversampling Technique (SMOTE) forms new minority class examples by interpolating between several minority class examples that lie together. This algorithm creates new instances of the minority class by creating convex combinations of neighboring instances. The main disadvantage is that it does not solve the problems associated with skewed class distribution. SMOTE tends to create a large number of noisy data points. Since clusters are not well defined, some majority class examples might be present in the minority class space and vice versa. This will cause overfitting.

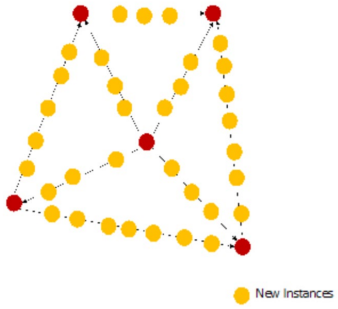

Figure 7. Adding samples to the minority class by using SMOTE

By using Tomek links for undersampling, the overfitting problem caused by the random undersampling method is mitigated. A Tomek link is a link between the two nearest neighbor examples from different classes. The example belonging to the majority class in a Tomek link is removed. This approach removes the noise or borderline examples to create well-defined clusters of majority and minority classes.

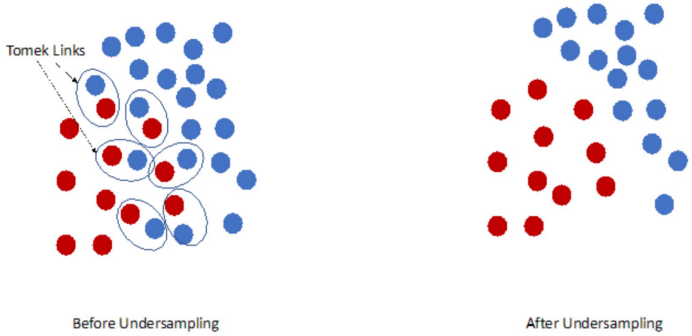

Figure 8. Balancing Dataset by using Tomek

In the SMOTE+Tomek Links approach, we first use oversampling using the SMOTE method and next, do the undersampling using Tomek Links. After oversampling by SMOTE, the class clusters may be overlapping with each other. Now, we do undersampling by using Tomek links. However, with the SMOTE+Tomek approach, we remove examples from both classes instead of from only majority class examples as in the Tomek approach. We chose to use the SMOTE+Tomek links approach after comparing several metrics against the SMOTE approach using a testing algorithm. Table 1 shows our results, averaging each metric for 4 different datasets.

Table 1. SMOTE vs SMOTE+Tomek resampling results (to 3 d.p.)

Resampling Method	Precision	Recall	F1	MCC	ROC_AUC	Gini
SMOTE	0.545	0.819	0.567	0.475	0.767	0.534
SMOTE+Tomek	0.781	0.833	0.803	0.678	0.848	0.698

2.3 Reject Option Classification (ROC)

We intend to utilize the Reject Option-based Classification technique, based on the decision theory for discrimination aware classification to mitigate bias after the classifier outputs predictions.
We rely on the assumption that most discrimination occurs when a model is least certain of the prediction i.e. around the decision boundary (classification threshold). Thus, by exploiting the low confidence region of a classifier for discrimination reduction and rejecting its predictions, we can reduce the bias in model predictions. For example, with a classification threshold of 0.5, if the model prediction is 0.81 or 0.1, we would consider the model certain of its prediction but for 0.51 or 0.49, the model is not certain about the chosen category. In ROC, for model predictions with the highest uncertainty around the decision boundary, when the favorable outcome is given to the privileged group or the unfavorable outcome is given to the unprivileged, we modify them. With this post-processing technique, developers would be able to mitigate bias without modifying the learning stage and so are not restricted by any specific learning algorithm.
To evaluate our hypothesis, we refer to the ROC technique applied to model predictions on the Adult Dataset[IBM] by Haniyeh Mahmoudian, TDS.

Table 2. ROC evaluation results

Models	Accuracy	Disparate Impact	Average Odds Difference
Predictions without mitigation	0.85	0.36	-0.35
Prediction with ROC	0.78	1.0	-0.05

Though the accuracy dropped by 0.07, this technique was able to substantially improve the DI score and reduce the Average Odds Difference almost to zero, indicating that ROC was able to successfully mitigate bias with a small trade-off in accuracy.

2.4 Benefits of Pre-Processing Techniques

Fairness is a highly complex issue that we cannot expect an algorithm to define on its own. Research has shown that training an algorithm to perform equally on all subsets of the population will not ensure fairness and will instead cripple accuracy. Adding extra objective functions can hurt model accuracy, causing a tradeoff. Therefore, it is better to keep the algorithm simpler and ensure the data is balanced—improving model performance and avoiding the tradeoff. Additionally, it's unreasonable to expect a model to perform well on cases for which it's seen few examples. Consequently, the best way to ensure good results is to improve the diversity of the data. From a commercial point of view, attempting to de-bias a model with engineering techniques is expensive and time-consuming; It's much cheaper and easier to train your models on unbiased data in the first place, freeing up your engineers to focus on applications.

3 Implicit Bias

Implicit Bias is an essential part of mitigating bias in AI. Since developers are involved in every step of creating an AI system, their own biases can enter the system in various ways. When determining the aim of the system at the very start, developers are prone to confirmation bias - failing to consider more nuanced systems and the full scope of the problem - especially if they are unaware of implicit bias. This can translate into prejudicial and/or measurement bias through biased or unbalanced training data. While developers may not be consciously creating a biased system; bias enters the system nonetheless. Developers themselves are not necessarily at fault for this; however, training them to recognize implicit biases can help reduce the bias entering these systems in the first place.

There are many kinds of implicit biases that can affect the development of an AI system. The most publicized kind is prejudicial bias against certain races, genders, or other groups. While this is an issue, there are other kinds of bias as well. For instance, confirmation bias causes developers to fail to see how systems could fail. In their desire for the system to succeed, they neglect its potential failures and shortcomings. Additionally, measurement bias fails to account for the differences between the model's training and its application in real life, thus creating a system that is unprepared for accurate decision-making in various circumstances. While bias can usually be fixed even after the system has been implemented, this needlessly causes thousands of people harm that could have easily been prevented, as well as reinforcing already-existing systemic biases in our society. By working to reduce implicit bias, we also work towards a more just society.

3.1 Guidebook

AI404's Guidebook to Mitigate Bias (https://bit.ly/3huXUHM) works towards removing or at least reducing the amount of implicit bias that enters an AI system at all. It first raises awareness about implicit bias and overcomes resistance to the idea, starting from the very bottom and building up to the creation of a fair system. It also shares the blame - both individual employees and companies have to take responsibility for mitigating bias on both the individual and company-wide levels. However, it is also nuanced, taking into account different kinds of bias and differences between companies. It takes a multi-layered approach that goes beyond simply advising training exercises, and provides several long-term changes to create a workplace in which mitigating bias and working towards equity are key concerns. And, although bias mitigation is often lauded as a cure-all, it must be used in tandem with other measures, like our dataset testing GUI, in order to effectively reduce the bias that enters AI Systems.

3.2 Explainability

As AI models are now being used in important aspects of people's lives such as medical diagnosis, crime forecasting in criminal justice systems, and loan approvals, the need to explain the decisions made by these AI systems arises. The more capable an AI is in explaining the reasoning behind its decisions, the more trustworthy and transparent the AI models become. According to Amit Paka, the co-founder and CPO of Fiddler Labs, explainable AI can improve several processes such as the AI-based credit lending model used by banks as shown in the table on the right. Paka said that "There are a number of inputs (like annual income, FICO score, etc.,) that are taken into account when determining the credit decision for a particular application. In a traditional environment without Fiddler, it's difficult or near impossible to say how and why each input influenced the outcome." Paka added that by using explainable AI, banks could now "attribute percentage influence of each input to the output. In this case, an example could be that the annual income influenced the output positively by 20% while the FICO score influenced it negatively by 15%." From Paka's statements above, it can be concluded that after applying AI Explainability into the credit lending model that the banks use, they can now identify which factors affected the decisions made by their AI models. These explanations are essential in a way that they let the human operators observe whether the models that they made are not reflecting bias. And in the case that the explanations do reflect bias, they can now prevent bias from occurring in the future by applying some modifications into their overall model.

One approach to implement explainability in AI is to look into the scope of the explanations, which could either be global or local. Global explainability refers to the ability of the model to explain how the overall algorithm functions (Guidotti et al., 2018) and how decisions are generally made based on a holistic view of its features and each of the learned components such as the parameters, weights, and structures. Global explanations give an overview of the AI system, how the AI learned, what training data were used, and the limitations to the model and use restrictions. On the other hand, local explainability refers to the ability of a model to explain how a specific decision was made such as what the main factors were considered to arrive at that decision. Providing explanations of how a decision was made largely depends on the context specifically on the target audience of these explanations. For instance, the end-users of these softwares need different explanations from those who are developers. To efficiently mitigate bias, the models need to look at a broader explanation that will give more insight on how the overall system works which will let the human operators or the developers identify when the AI systems will likely perform well or badly, or when the system will likely show bias in their decision. In this case, the global explainability model best fits the problem of mitigating bias. In applying for credit loans, several cases have been reported wherein certain minorities and ethnolinguistic groups were penalized by AI systems through their automated recommendation and decision-making. By applying the global explainability model to that AI system, it can now have an overall understanding of how they make their decisions. Hence, the cause of the bias to those minorities can now be identified and avoided in the future. With that, the developers can adjust and modify their algorithm to achieve a better and fairer system.

4 Regulations & Report Logistics

Bias in AI poses great risks especially when it negatively affects people and violates their rights. When this happens, statutory, contractual, and common laws may be violated which will make the companies and organizations liable for their AI systems. Although AI is a relatively new emerging field, many regulations have been enacted by the Federal Trade Commission (FTC), the European Court of Human Rights (ECHR), and other law-enacting bodies in other countries to protect the people from AI's unfair decisions.

Specific Regulations -

- Section 5 of the FTC Act - The FTC Act prohibits unfair or deceptive practices. That would include the sale or use of – for example – racially biased algorithms.
- Fair Credit Reporting Act - The FCRA comes into play in certain circumstances where an algorithm is used to deny people employment, housing, credit, insurance, or other benefits.
- Equal Credit Opportunity Act. - The ECOA makes it illegal for a company to use a biased algorithm that results in credit discrimination on the basis of race, color, religion, national origin, sex, marital status, age, or because a person receives public assistance.

We propose the creation of a B2C form involving government intervention that affected consumers can fill to report bias. If the predictions made are discriminatory, due actions can be taken.

5 Proposed Improvements

Edited Nearest Neighbor (ENN) is an undersampling technique where the nearest neighbors of each of the majority class are estimated. If the nearest neighbors misclassify a particular instance of the majority class, then that instance gets deleted. In this approach, the oversampling is done by using SMOTE, and undersampling is done by using ENN. This results in a more clear and concise class separation. However, SMOTE+ENN is incredibly aggressive at downsampling datasets, and the initial datasets we had used to test our algorithms varied with results with each technique, and the implementation of SMOTE+ENN required much more labor. Due to time constraints, we used the SMOTE+TOMEK approach instead. However, a SMOTE+ENN solution would have better performance, and this is a point of future improvement. Generative adversarial networks (GANs) have been recently increasing in popularity, and we have also attempted to utilize them in our GUI. However, the complexity of GAN models was too high to work within our scenario. Further research in GANs would be a huge step forward in the future.

6. Conclusion

In this paper we have outlined the structure and benefits of our proposed algorithmic framework to counter bias in AI/ML systems composing of a general guidebook, an automated de-biasing algorithm, SMOTE-Tomek, and a Reject Option Classification algorithm. As reported previously, we have devised a strategy incorporating a novel decentralized database to retrieve additional data used for representative datasets. Consequently, this framework

represents a powerful tool for teams of all sizes and of all skill levels to address the potential discriminatory predictions by its artificial intelligence/machine learning solution.
This study is an important step towards a user-friendly, accessible approach for common but severe issues in AI solutions, in this case bias in artificial intelligence, in a world in which AI is becoming increasingly important. On a wider level, research and practical implementation is also needed to determine the solution's suitability in a large-scale industry or real-world scenario. However, the prospect of AI being able to transform virtually all sectors and the formidable danger of discriminatory towards specific population groups serves as a steady incentive for further research.

References

Batista, Gustavo. "Balancing Training Data for Automated Annotation of Keywords: A Case Study." ... of the Second Brazilian Workshop on ... Accessed August 15, 2021. https://www.academia.edu/34702950/Balancing_Training_Data_for_Automated_Annotation_of_Keywords_ a_Case_Study.

"Bias in Data-Driven Artificial Intelligence Systems-An Introductory Survey." ResearchGate. Accessed August 15, 2021. https://www.researchgate.net/publication/338998132_Bias_in_data-driven_artificial_intelligence_systems-An_introductory_survey.

Chiappa, Silvia. "Path-Specific Counterfactual Fairness." Proceedings of the AAAI Conference on Artificial Intelligence. Accessed August 15, 2021. https://ojs.aaai.org//index.php/AAAI/article/view/4777.

Damon Civin, ARM. "Explainable Ai Could Reduce the Impact of Biased Algorithms." VentureBeat. VentureBeat, May 21, 2018. https://venturebeat.com/2018/05/21/explainable-ai-could-reduce-the-impact-of-biased- algorithms/.

"Fairness Metrics: A Comparative Analysis." IEEE Xplore. Accessed August 15, 2021. https://ieeexplore.ieee.org/document/9378025/authors#authors.

Gerards, Janneke, and Frederik Zuiderveen Borgesius. "Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence." by Janneke Gerards, Frederik Zuiderveen Borgesius :: SSRN, January 7, 2021. https://doi.org/10.2139/ssrn.3723873.

"If You're De-Biasing the Model, It's Too Late." Scale. Accessed August 15, 2021. https://scale.com/blog/if-youre- de-biasing-the-model-its-too-late.

Lee, Nicol Turner. "Detecting Racial Bias in Algorithms and Machine Learning." Journal of Information, Communication and Ethics in Society. Emerald Publishing Limited, August 13, 2018. https://doi.org/10.1108/jices-06-2018-0056.

Oneto, Luca, and Silvia Chiappa. "Fairness in Machine Learning." arXiv.org, December 31, 2020. https://arxiv.org/abs/2012.15816.

Prasuna. "Data Bias: What Is Data Bias: How to Reduce Bias." Analytics Vidhya, March 30, 2021. https://www.analyticsvidhya.com/blog/2021/03/fighting-data-bias-everyones-responsibility/.

Rao, Anand, and Ilana Golbin. "What Is Fair When It Comes to Ai Bias?" strategy+business, April 12, 2019. https://www.strategy-business.com/article/What-is-fair-when-it-comes-to-AI-bias.

"Root out Bias at Every Stage of YOUR AI-DEVELOPMENT PROCESS." Harvard Business Review, October 30, 2020. https://hbr.org/2020/10/root-out-bias-at-every-stage-of-your-ai-development-process.

Ross, Andrew Slavin, Michael C. Hughes, and Finale Doshi-Velez. "Right for the Right Reasons: Training Differentiable Models by Constraining Their Explanations." IJCAI, January 1, 1970. https://www.ijcai.org/proceedings/2017/371.

Silbermann, Thierry. "A Look at the Methods to Detect and Try to Remove Bias in Machine Learning Models." InfoQ. InfoQ, September 26, 2019. https://www.infoq.com/presentations/bias-ml-sao-paulo-2019/.

Stefanie Koperniak | Institute for Data. "Artificial Data Give the Same Results as Real Data - without Compromising Privacy." MIT News | Massachusetts Institute of Technology. Accessed August

15, 2021. https://news.mit.edu/2017/artificial-data-give-same-results-as-real-data-0303.
"A Survey on Data Collection for Machine Learning: A Big Data - Ai Integration Perspective." IEEE Xplore. Accessed August 15, 2021. https://ieeexplore.ieee.org/document/8862913.
Team, The Salesforce Einstein. "Trusted Ai: Finding Bias in Your Data." Medium. Salesforce Einstein Platform, May 28, 2020. https://medium.com/salesforce-einstein-platform/ep-trusted-ai-blog-finding-bias-in-your-data- 3df0f057fd19.

Biography

Tanish Kelkar is a junior high school student from Maharashtra, India. He is proficient in Python, C, C++, Java, JavaScript, SQL, and HTML, excelling in utilizing Linux-based IDE's and bash scripts to automate tasks. He has researched projects in fields related to 3D Designing and simulation such as Robots, solar dehydrators, solar chargers, Polymer-coated Photo-Voltaics and Solar Trees. Currently, he works on modeling an SQL Database containing OPD Patients' Appointment data over the last 12 years with over 4000 data points into a Snowflake Schema Data Warehouse, with the aim of devising a greedy algorithm to optimize outpatient scheduling in multispecialty hospitals using the modeled data.

Ashug Gurijala is a junior high school student from Telangana, India. He has received distinction in IGCSE and is currently pursuing an IBDP diploma, planning to pursue Computer Science in the future. He has programming experience in Python and web/app development, utilizing this experience for research. He is currently conducting research in single cell RNA sequencing analysis.

Shizhe He is a high school junior of the TUMKolleg program in Munich, Germany. He has both practical experience through numerous projects and research experience in the fields of machine learning and computer vision in computer science. Additionally, he is currently working on a thesis on MRI reconstruction using deep learning at the Lab for AI in Medicine at the Technical University of Munich.

Avni Gupta is a high school student from Connecticut, USA. She is experienced in programming, especially in web and app development, using HTML/CSS as well as Java. She also has experience doing bioinformatics research with R, unix, and supercomputers, and has previously won the Congressional App Challenge. In the future, she plans to work on computational biology and cybersecurity.

KIAN ANDREW BUSICO is a high school student from the Philippine Science High School, Southern Mindanao Campus. Davao City, Philippines. He is experienced in mathematics and computer science, currently researching cybersecurity, data science, forensic science, mathematics, and technology.