Supporting Documentation Template

Submission Advice for Meeting Judging Requirements

Supporting documentation, including any tables, charts or graphics, **must not exceed the maximum page count (10 pages)**. The use of tables, charts, and other illustrative graphics is encouraged to the extent that it enhances readability and understanding.

Please note the lengths provided for each section below are recommendations. The length of sections can be adjusted; however, the total page count must not exceed 10 pages, excluding title page and references. Anything after 10 pages will be ignored.

Refer frequently to the problem statement and judging criteria as you develop your solution to ensure your solution remains focused!

Consider the overall objective of this effort - to improve healthcare by detecting and mitigating bias in AI/ML algorithms. You need to explain the value of your solution to healthcare providers and patients so they are comfortable using AI to make clinical decisions and following those healthcare recommendations

Title Page (does not count toward page limit)

Include:

- Solution Name/Team Name
- Link to Video
- List of team members
- Contact information (email and phone number) for team's point of contact
- Abstract (200 words, does not count toward page limit)

GitHub code (less than 1 page)

Although you submitted your code in GitHub, please Include a link here for easy reference. We suggest including the actual output and/or visualizations from your bias detection tool and anything else you need to submit to demonstrate input/output.

https://github.com/equialgo/fairness-in-ml/compare/master...pamegup:fairness-in-ml:Bias-Detect-and-Remove-Tool?diff=split

Methodology Overview (less than 1 page)

Describe the approach you took for disparity measurement and bias mitigation. Do not explain AI/ ML, modeling, the ethical concerns bias poses to healthcare, or the historical, cultural or demographic reasons for this challenge. Focus on the analysis and design choices you made to identify, mitigate and monitor bias in healthcare models. While this challenge requires you to use healthcare models from healthcare data, that prerequisite effort will not be judged. You could include written explanation, math, visualization, and highly technical components, but keep in mind that effective documentation also contains information useful to an interdisciplinary audience.

Your submission will be evaluated by the judges based on how well this section answers these questions:

- What metrics did you choose to identify different types of bias and why? Which types of bias did you identify and how?
- How does your bias detection tool measure retrospective and prospective

bias? ● How does your tool detect “latent” bias or drift?

- What creative or novel approaches did you incorporate in your bias detection tool?
- If the tool creatively departs from existing practices and standards, how well-justified are those departures?
- What was your inspiration for this methodology?

Although AI/ML algorithms offer promise for clinical decision making, that potential has yet to be fully realized in healthcare.

Even well-designed AI/ML algorithms and models can become inaccurate or unreliable over time due to various factors;

Changes in data distribution, subtle shifts in the data, real world interactions, user behavior, and shifts in data capture and management practices can have repercussions for model performance.

These subtle shifts over time can cause degradation of the predictive capability of an algorithm, which can effectively negate the benefits of these types of systems in the clinic. Accurate monitoring of an algorithm’s behavior and the ability to flag material drifts in performance may enable timely adjustments that ensure the model’s predictions remain accurate, fair, and unbiased over time. In this way, degradation of the predictive capability of the algorithm when applied in the real world may be prevented.

As AI/ML algorithms are increasingly utilized in healthcare systems, accuracy, generalizability, and avoidance of bias and drift appropriately come to the forefront. Bias can primarily surface in the form of predictive bias—algorithmic inaccuracies in producing estimates that significantly differ from the underlying truth; and/or social bias, or latent bias.

1. **How do you identify predictive and social bias?**

Predictive Bias: Algorithmic inaccuracies in producing estimates that significantly differ from the underlying truth.

Social Bias: Systemic inequities in care delivery leading to suboptimal health outcomes for certain populations.

We aim to identify the predictive and social biases of gender, ethnicity and ICU referrals with reference to COVID-19 and de-bias these metrics in the model outputs for health practitioners to utilise for de-biasing their own datasets. We will ensure fairness metrics are built to capture model behaviour on different protected characteristics. e.g. Difference in Equality of Opportunity (DEO) / BA [bias amplification] ["Mitigating Bias in AI Using Debias-GAN" p. 8]

Value Proposition (less than 1 page)

Introduce what is unique and innovative about the approach you chose and demonstrate its value to the healthcare industry. It will probably be helpful to highlight **how the output of your tool communicates what bias it found and how users can take corrective action and guard against future bias.** Without spending too much time "admiring the problem" of bias in healthcare or the importance of ensuring models aren't trained on biased data, explain the value your tool brings to the healthcare ecosystem by focusing on **how it creates transparency and explainability that foster trust and confidence in AI/ML.**

Your submission will be evaluated by the judges based on how well this section answers these questions:

- **What types of biases are you addressing and how accurately does your tool identify bias within the algorithm?**
- What appropriate **metrics are proposed to identify the bias?**
- Did the t**ool identify the type of bias** t inherent in the algorithm?
- Is the bias **predictive and/or social** and how did the tool identify these biases?

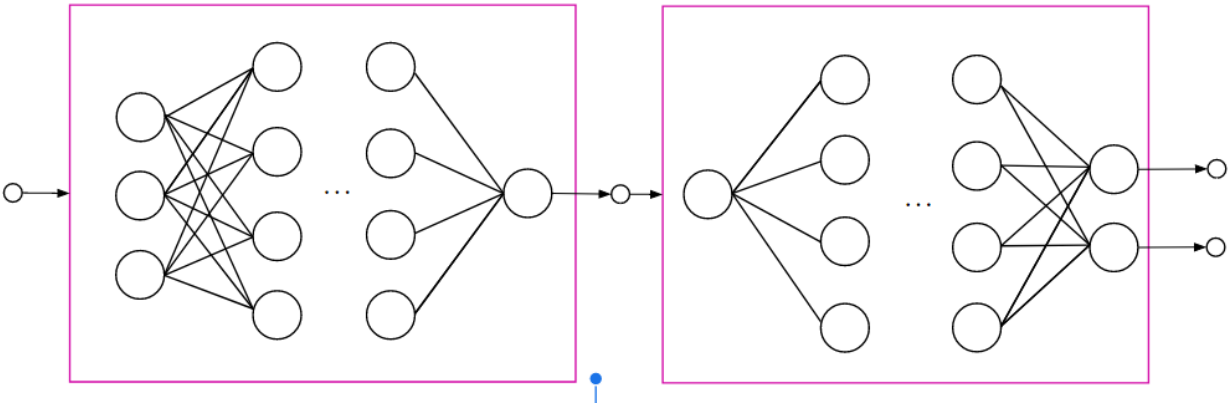
We identified the biases present in synthetic datasets by analysing patient ICU referrals by gender and ethnicity for COVID-19 using seaborn. Once identified By building and training debiased LMs with generative adversarial networks (GAN) through reinforcement learning (RL), The debias-GAN is able to improve the fairness metrics of the classifier by up to seven times while maintaining classification performance.

Using python and Jupyter notebooks we have demonstrated the removal of the bias present in our initial analysis, we also observed improved fairness metrics [p. 15] based on our biases identified.

It will bring value to the healthcare sector by ensuring medical practitioners have the ability to understand the numbers of ICU referrals for COVID-19 and gender as well as ethnicity metrics and be able to trust and be confident in the use of the model as we focus on transparency and explainability in building, training and deploying our ML application.

[p17]Our study provides a solid example of utilizing AI to fix biased input data that correlates conversational tweets and user ethnicity. Likewise, the same approach can also be applied in real life tasks to combat biases in AI, especially those that result from biased training data

GAN [Generative Adversarial Network] using classifier and adversarial networks.



Our value proposition and advantage is in the data and model we have developed.

Our data strategy

Our data strategy is two fold, ensuring data security, privacy, integrity, quality, regulatory compliance and governance as well as improving our competitive position. Our data strategy will include the core activities of optimising data extraction, standardisation and storage as well as access. It will also ensure optimisation of data analytics and enrichment.

Our Digital value chain

Our digital value chain involves considering data, data services, exploitation, decision services and organisational performance; whilst defining organisation roles and leadership, process and procedure and the tools and technology required for the application to be deployed. Fundamental to this is good governance, regulatory compliance and risk management.

We took the Synthia MASS 10k COVID-19 dataset, which does not contain any PII that we are exposing and is synthetic.

We comply with the following data standards - ISO 8000.

Healthcare Scenario (1 page)

Describe how your tool meets the functional challenge objectives by stepping through how your tool would be used in a healthcare setting to address bias. Explain the medical context

and/or specific decision or prediction (diagnosis, recommendation for treatment, prognosis) that your tool would address, and how your tool could be extended to cover other clinical decision scenarios.

The tool would be used in a healthcare setting to build trust and confidence in the analysis of patient data with reference to ICU patients, gender and ethnicity by removing initial bias indicators to reduce the unintended consequences of Bias in AI and the tendency to assume bias when referring to particular ethic groups or genders when recommending care programs or diagnosing healthcare concerns. [p.4 'Mitigating Bias in AI Using Debias-GAN']

Your submission will be evaluated by the judges based on how well this section answers these questions:

2. **How do you account for "latent" bias where social or statistical biases happen over time due to the complexities of healthcare processes?**

Biases that emerge from adaptive AI-based algorithms after they are deployed can be best described as "latent biases," (ie, biases waiting to happen). latent errors in complex systems are understood as 'errors waiting to happen'. We account for this bias by analysing our model performance over time and providing reports on this.

Our tool will address latent bias such as social and statistical bias over time by monitoring the model's training and results based on the parameters it is trained upon and providing reported results.

3. **Where does bias occur and how do we provide a path forward for follow-up investigations?**

Bias occurs in ML models at the early stages of data analysis and our model is designed to remove the identified social biases of gender, ethnicity and ICU referrals from COVID-19. Then targeted language model (LM) generates realistic but ethnicity and gender -oblivious outputs. We trained such debiased LMs with generative adversarial networks (GAN) through reinforcement learning (RL)

4. **. How do you account for consistent evaluation and assessments of the algorithm over time and for all patient populations?**

We will account for consistent evaluation and assessment of the algorithm over time by working out the algorithm's runtime as a function of the problem-specific instance features and parameters and optimising this by ensuring the model can run successfully within a given captime and by analysing performance over time. By de-biasing we lower risk level in across all patients.

- What recommendations did your tool suggest to identify the underlying root(s) of the biases detected? Who would be responsible for investigating cause and remediation options? How easily could the tool be implemented to perform this task?

Internally engineers and ML experts with knowledge of the risks and ramifications of bias emerging in AI and data trained on models would be responsible for investigating cause and remediation options to intervene to ensure balance and reduction in latent biases. The tool can be implemented easily with access to GPU and local python libraries and notebooks.

- How will your bias detection tool, when implemented as described here, improve healthcare over time and for all patient populations?

Over time it will reduce the need to re-train and evaluate data outputs of ML Models by building trust, transparency and confidence overall in the tool built, which will reduce risks and unintended consequences of bias in diagnosis and care plan evaluations. one of the many sources of the bias in AI, and our approach could be further applied to real life tasks that are currently plagued by inequities in AI applications.

Operational Requirements (less than 1 page)

Specify the technical and operational requirements for your tool. **You may provide a configuration diagram or listing of OS, memory (RAM), disk space, CPU/GPU used, and any required environment configurations required to execute the code.**

Operational Requirements are as follows:

Google Colab Pro - CPU details for Google Colab : 2vCPU @ 2.2GHz

13GB RAM

GPU: 1xTesla K80 , compute 3.7, having 2496 CUDA cores , 12GB GDDR5 VRAM

CPU: 1xsingle core hyper threaded Xeon Processors @2.3Ghz i.e(1 core, 2 threads)

Your submission will be evaluated by the judges based on how well this section answers these questions:

- Can this tool be deployed widely and if so, is an SOP and/or code etc. provided for implementation at other locations?
- Are there any dependencies on vendors or proprietary information exchange standards that would limit the tool's impact?
- What is your architectural design and how do you ensure it is **technology-agnostic?**
- Do you acknowledge that the tool will be distributed under the **BSD 3 license?**

Sustainability Plan (1 page)

Describe how you envision a healthcare organization sustaining your bias detection tool. List the responsibilities and activities required across the different stakeholder groups who must be involved in implementing your tool (e.g. technologists, informaticists, doctors, practitioners, administrators, etc.). Elaborate on the ongoing level of effort to monitor and mitigate bias on a continual effort, including tool maintenance and retrospective analysis to guard against drift.

Your submission will be evaluated by the judges based on how well this section answers these questions:

- Does your plan take into account or involve the multidisciplinary team and expertise needed for implementation in real-world settings?

Our sustainability plan would involve the input of a multidisciplinary team with real world experience in a health care setting to ensure it is sustainable and performing optimally.

- Does your tool accurately calculate retrospective (measuring performance of the clinical decision model when developed) and prospective (measuring performance continuity/in real world application) metrics?

Our model aims to improve runtime predictions of AI planning methods and used the resulting predictions to compute captimes that maximize a given utility function. Gagliolo et al. [28,27]

- Can the tool be used in the future without significant upkeep costs?

Our solution is industry and vertical agnostic and our aim is to create a library routine that any developer can provide input what bias is that they wish to detect and make it routine and build a model to enable them to de-bias their own data and it will reduce over cost and resource does not have to be spent to figure out their own solutions, routines will be efficient and no need to re-create the code.

The solution will make ICU admissions in hospitals fair and not based on race and gender, as hospitals are highly regulated organisations under HIPPA compliance requirements and this will require a particular data strategy.

If deployed with an efficient runtime and deployed on GPUs on cloud platforms that are cost efficient the

up keep costs will be minimal. this must be monitored and periodically evaluated.

Generalizability Plan (1 page)

Understanding that you chose to develop and test your tool in a particular technical environment which may not match the infrastructure and resources available in a healthcare setting, explain why your tool is **portable and extensible.** Describe how your tool can be easily extended to address additional healthcare scenarios and clinical decisions, or how the output from your tool can be shared with **patients and patient advocacy groups** to nurture trust in the models that your tool is monitoring. Highlight the **design choices** you made to optimize the usability of your bias detection tool's output and how **actionable and helpful it is for the future users** (data scientists, healthcare providers, etc.). This plan should substantiate how well your tool meets the broader, social purpose of this challenge.

Your submission will be evaluated by the judges based on how well this section answers these questions:

- **How impactful is the tool?** Who can use this tool and how likely are they to use it? How easily can the target user run it on an ongoing basis to monitor and continually improve bias mitigation practices? How easily could your tool be applied to other and broader types of clinical decisions, i.e. predictions, diagnoses, or treatment recommendations?

- **Can this tool be deployed widely and if so,** is an SOP and/or code etc. provided for implementation at other locations? How easily could your tool be leveraged in other clinical disciplines or with other types of health information, e.g. cardiology, oncology, obstetrics, etc.?

- The tool can be leveraged in any clinical discipline whereby there is a need to de-bias patient data in a healthcare setting, such as A&E, ICU, cardiology

- **What are the next steps for this tool? Can this tool be implemented as is or does it need additional support?** *It can be implemented with appropriate python libraries and notebooks.*

- **Did the tool suggest follow-up investigations to identify the underlying root(s) of the biases detected? How easily could the tool be implemented to perform this task?** *We completed an initial analysis of the datasets in Seaborn for analysis of existence of bias and found correlation between gender and ethnicity and numbers of ICU cases, it did not suggest follow up investigations for other root causes.*

- **Does the tool seem likely to improve health equity** (fairness) and trust in AI/ML in healthcare settings?

- **How easily could your tool be applied to other and broader types of clinical decisions,** i.e. predictions, diagnoses, or treatment recommendations?

It may be deployed in support of treatment diagnosis and care settings

- **How easily could your tool be leveraged in other clinical discipline**s or with other types of health information, e.g. cardiology, oncology, obstetrics, etc.?

The tool can be leveraged in any clinical discipline whereb there is a need to de-bias patient data in a healthcare setting, such as A&E, ICU, cardiology,

Implementation Requirements (1-2 pages)

Describe the system and human resources that would be required to scale usage of this tool so the recommendations are followed at multiple locations. Consider if the tool would be used by professional organizations, research or academic institutions, or in healthcare delivery systems.

Your submission will be evaluated by the judges based on how well this section answers these questions:

- Do the implementation strategy and SOPs take into account or involve the **multidisciplinary team** and expertise needed for implementation in real-world settings?
- Can this tool be deployed widely and if so, is an SOP and/or code etc. provided for implementation at other locations?

● Describe the human resources needed to implement in a real-world setting? ● How do you measure success of the implementation?

This would require a Data Scientist or ML Engineer with experience of training/ adjusting and maintaining GAN Algorithms to ensure controls and accuracy over time [p.6 'Mitigating Bias in AI using Debias GAN']

to what extent the tools developed improves trust and confidence in ML/AI tools and usage across healthcare by monitoring adoption over time will determine implementation success.

Lessons Learned (less than 1 page)

Describe what challenges you ran into and how you overcame them. And please share the lessons you learned from this challenge experience. Also you may optionally add any thoughts or recommendations for NCATS or other organizations on how they can build on this challenge to further address the broader issue of improving the **value and transparency of AI/ML for clinical decisions.**

Your submission will be evaluated by the judges based on how well this section answers these questions:

● How will your tool, next year or three years from now, improve health equity (fairness) and trust in AI/ML in healthcare settings?

● Are there other ways your tool can be used, perhaps in a research or government setting, to improve healthcare outcomes? or address other social inequities?

The first challenge was to identify the most appropriate dataset that would provide the most complete data on gender, ethnicity and ICU admission. We initially analysed Columbia public data set for sample synthetic datasets.

A need to better balance the datasets by swapping gender specific terms like men, women, name anonymisation and gender tagging with NMT at the beginning of a datapoint, to reduce bias further; as well as adjusting the algorithm further by using a discriminator to identify gender in adversrial learning. ["Mitigating Bias in AI Using Debias-GAN",*p. 6]*

Build a SeqGAN model to train the model in reinforcement learning for our discrete values [tokensisation] with monte carlo search [p. 6] to improve the tool's de-bias capabilities. & leveraged a reinforcement learning (RL) framework with a generator with a penalty for generating biased outputs as part of adversarial training and employ the "debiased" generator to synthesize a balanced dataset as the input for the downstream text classifier.

improve the process of tagging fields / datasets to improve efficiencies.

In future iterations we could ix synthetic data of varying sequence length with real data as input to the downstream natural language process models

A more complicated model could be leveraged, such as GPT 2 or 3 by open AI. The goal is to fine tune the last few layers of the model rather than train it from scratch

Within the debias-GAN, a Monte Carlo tree search and value network can be implemented instead of the current rollout strategy to improve action decision making

A debias-GAN framework will be popular and widely applied in a broad range of use cases, and truly improve AI decision making and promote diversity and inclusion["Mitigating Bias in AI Using Debias-GAN",*p.17]*

References

www.towardsdatascience.com

Synthia COVID-19 dataset

What's the hardware spec for Google Colaboratory? - Stack Overflow

"Mitigating Bias in AI Using Debias-GAN", WWT Artificial Intelligence Research & Development, October | 2021]

Algorithm runtime prediction: Methods & evaluation | Elsevier Enhanced Reader

Questions? Post them in Slack (#main channel) or email expeditionhacks@blueclarity.io.