ECSE 557

Intro to Ethics of Intelligent Systems

Final Report

CHIKHAOUI Hamza 260912960

GUMUS Okyanus 260900481

LAMBRIANOS STAPPAS Anastasios 261009936

Winter 2024

Introduction

Within the past few years, artificial intelligence (AI) and predictive models have emerged as powerful tools in various domains. They can process significant amounts of data and extract trends that were too complicated for previous computation methods. In healthcare, specifically, AI has the potential to offer quick diagnosis of diseases based on patients' medical histories. However, since predictive models demonstrate a significant dependence on the training data, a model bias stemming from the training data is a cause for concern.

The ramifications of gender bias in AI medical diagnosis are extensive. Misdiagnosis or underdiagnosis of certain conditions in specific gender groups can have profound implications for patient health outcomes, leading to delayed treatment, inappropriate interventions, or even worsening of medical conditions. Furthermore, perpetuating gender biases in healthcare through the use of biased AI models have the potential to reinforce existing disparities and inequalities.

The nature of this bias can stem from various factors, ranging from the specific nature of a disease, causing it to be more prominent in certain demographic groups, to physiological and environmental factors. In this project’s case, gender will be the primary consideration. Gender bias in AI models manifests themselves as systemic inaccuracies or disparities based on the gender of the individuals. In the context of medical diagnosis, gender bias can severely impact the quality of the diagnosis process and the care received by the patients.

The root cause of gender bias within an AI model can be attributed to different sources, such as algorithmic bias, inequalities within the healthcare system, or data collection bias. This data collection bias can arise from various underlying causes, such as societal norms, cultural perceptions, and healthcare delivery. These root causes can manifest as different healthcare-seeking behaviors between genders and affect the presentation and diagnosis of medical conditions, which in turn will affect the training data of a specific demographic group for a specific disease. Hence, it must be mentioned that gender bias in medical data also depends on societal and cultural norms, but these are out of this project's scope.

In this project, the data collection bias will be the main focus where we will be considering the impact of gender underrepresentation within the training data. Based on the natural prevalence of a particular disease across genders and a specified gender distribution within a dataset, the objective of this project is to devise a measure to assess the anticipated bias in predictions of the model prior to the development of the model.

Background

Bias, in our project, is referred to as any systematic error made by the model based on the gender of the samples. Any systematic error originating from an unfair model causes prejudice to one gender group over the other. There are many possible sources of unfairness in machine learning (ML), as depicted in the following figure.

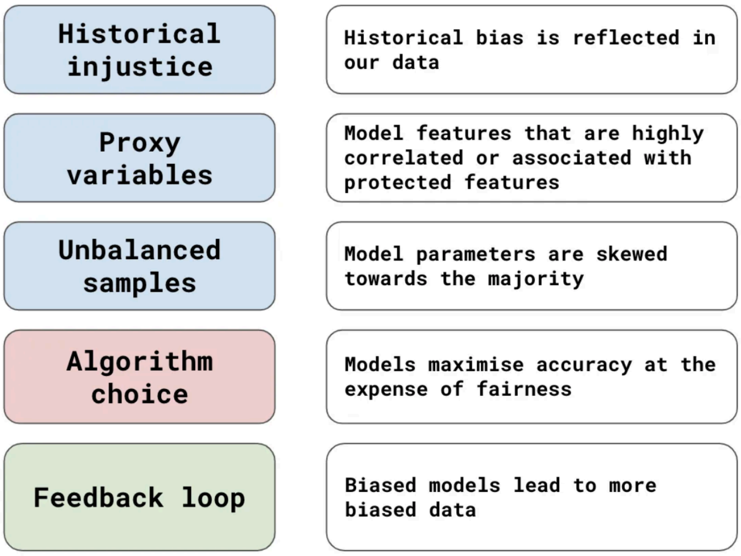

Historical injustice
Historical bias is reflected in our data
Proxy variables
Model features that are highly correlated or associated with protected features
Unbalanced samples
Model parameters are skewed towards the majority
Algorithm choice
Models maximise accuracy at the expense of fairness
Feedback loop
Biased models lead to more biased data

Figure 1 : Overview of sources of unfair predictions in ML

As illustrated in Figure 1, there are many ways in which unfair predictions can occur in ML. Main focus in our project is mainly the balancing of samples. To achieve this, an investigation into how different gender splits affect the overall bias of the model was conducted. Two datasets were implemented, with one being the prevalence of Alzheimer's disease in male and female patients and the other on the prediction of strokes. Both datasets can be found in [1, 2].

Preliminary research on these diseases' prevalence in both populations is essential in this context to estimate bias. Two important facts in this aspect is the impact of gender on Alzheimer's. Almost two-thirds of Americans living with Alzheimer’s are women, and women in their 60s are more than twice as likely to develop Alzheimer’s disease over the rest of their lives as they are to develop breast cancer [3]. Relating to the social context of Alzheimer’s, one of the primary reasons more women are affected by the disease than men is their longer average lifespan, since age is the most significant risk factor for Alzheimer’s. Furthermore, research based in Sweden showed that starting around age 80, women have a higher likelihood of being diagnosed compared to men [4]. Research based in Taiwan also displayed that over seven years, women had a higher chance of developing the disease than men, while in Europe, it’s estimated that around 13 women out of 1000 develop Alzheimer’s annually, as opposed to 7 out of 1000 men [4]. In terms of age-specific incidence rates, between the ages of 65 and 74, the incidence rate is 4 per 1000 individuals, while between 75 and 84, this rate jumps to 32 per 1000. There is a significant increase to 76 per 1000 for those 85 and older.

The main literary reference regarding statistics on stroke for men and women comes from “The Impact of Sex and Gender on Stroke” by Rexrode et al. (2022) [5]. From their analysis, it can be observed that women face a disproportionate burden of stroke mortality and disability. In the United States in 2019, stroke was the third leading cause of death in women compared to fifth in men. Women accounted for 57.1% of stroke deaths in the same year, with stroke partaking 6.2% of all female deaths and 4.4% of all male deaths. On an annual level, there were 55,000 more fatal strokes in women than men. Notably, the lifetime risk of stroke is slightly higher in women at 25.1%, while in men, it is 24.7% from age 25 onwards. It is also worthy of mentioning the social context of stroke as women typically experience stroke at an older age than men, normally living alone or widowed and having a higher degree of disability at the time.

Based on the previously mentioned cases, women are more prone to develop certain diseases. However, that might not be the case for the datasets, as regional, demographic, and other socioeconomic factors can play an important role in the diversity of the dataset, as previously mentioned in the introduction.

Based on previous literature, most of the investigations of ML diagnosis bias were based on the classification accuracy of models of COVID-19 with varying demographic data (including but not limited to race and gender) [6, 7]. However, extrapolating findings from COVID-19 studies to other diseases like Alzheimer's may not be straightforward due to variations in disease characteristics, demographics, and risk factors. Therefore, it becomes imperative to conduct dedicated solutions to bias within the context of Alzheimer's to ensure the accuracy and fairness of our models. This entails not only understanding the specific biases that may arise due to disease-specific attributes, but also implementing effective strategies to mitigate them.

In addition to the unique challenges posed by disease-specific attributes in Alzheimer's, it's essential to acknowledge the nature of bias, which extends beyond demographic factors like gender. Factors such as socioeconomic status, geographic location, access to healthcare, and cultural differences can also influence the prevalence and presentation of Alzheimer's disease.

Although effort was made to accommodate the impact of non-demographic factors, the main emphasis was placed on gender due to time and resource constraints. While gender bias is significant, overlooking the influence of socioeconomic status, geographic disparities, and cultural nuances could lead to incomplete assessments and ineffective mitigation methods. By addressing bias from a broader perspective, we can enhance the relevance and applicability of our findings for future work, ultimately improving equity in healthcare and automated disease diagnosis.

Methods/Design

a) Design decisions made to date, process followed, and any social/policy implications related to the design are clearly articulated.

There are multiple social implications related to our design. First of all, we will strive for our design to be implementable in a real life setting where the disease detection would have direct repercussions on the patients' treatment. Such an implementation should ensure fairness between all groups represented in our dataset, whose treatment may be impacted. It is therefore a social policy of ours to ensure fairness in our design to prevent any discrimination based on the gender of the patients.

b) Translation of normative policies, values, and other considerations are clearly articulated in the design/implementation choices made by the team (e.g., how X has been translated into the software architecture, or Y value has been considered in choosing Z parameter).

Because our design fits in the broad domain of disease detection, our policies are heavily influenced by the possible usages of our model. To ensure our model would be compatible with a real-world implementation, we have decided to prioritize the detection of True Positive samples and the reduction of false negative samples to a maximum when iterating over different gender splits.

Furthermore, as a normative policy, we will strive for equal opportunity between the expected privileged group (men) and unprivileged group (women). That is, we will attempt to mitigate our biases to achieve equal odds as a guiding fairness principle.

This will practically be embodied in our design in two different ways. The first way we will enforce equalized odds on our dataset will be the use of the AIF360 libraries. We will iterate over preprocessing, in-processing and post-processing techniques to identify which one would be the most applicable to our model.

The second way in which equalized odds will guide our design is that we will aim to settle for the gender split of our dataset that achieves the most equalized odds for the unprivileged and the privileged groups. If we name as "k" the proportion of female samples in our modified dataset, we will choose k such that the difference in True Positive Rate and False Positive Rate in both populations is the closest to 0.

c) Expected results upon completion of the project (e.g., how will the team assess the performance of the system, what the prototype will be able to do) is/are clear.

Moving forward in the project, we expect to take a deeper look into the impact of gender ratios within disease classification through experimentation and the implementation of techniques that handle sample disparity issues in machine learning. By building upon those explorations, we aim to attempt to figure out a quantifiable measure of evaluating the degree of bias in machine learning models through the gender distribution of a dataset about the disease itself.

In doing so we aim to investigate computational approaches in measuring such discrepancies as well as numerically through different manual approaches. We hope that by navigating this gray area we will be able to develop a methodology for predicting degrees of gender bias in disease classification models. To ensure that our approach is appropriate we will be introducing numerous metrics, as discussed in class, such as equal opportunity difference and aiming to minimize it whilst respecting the literature that analyzes the diseases we are evaluating. In doing so we aim to develop a fair and robust framework that ensures that gender bias in disease classification is minimized. Essentially, our goal is to develop a bias index that aggregates various disparity measures into a singular value through combining different disparities, fairness-aware learning algorithms that will allow us to quantify the reduction of bias and counterfactual analysis that will estimate how dataset modifications between gender ratios will affect the model's outcomes.

Although an optimistic endeavor we aim to attempt the successful completion of this project and take a step forward in quantifying gender bias in disease diagnosis.

Results:

a) Clear Outline Method

This project has been subdivided into three main subcomponents. In the first part of the project, we have done an extensive statistical analysis focused on the Alzheimer dataset to study the prevalence of the disease across a series of features. In the second part of the project, we have decided to assess our model against normative outcomes. To do so, we have decided to split our dataset into two groups: male and female. We have then decided to build 2 separate models to predict the presence or not of Alzheimer's disease for each group, and we have the study the most important prediction features in each group (i.e. the features that are the most statistically relevant in predicting the presence or not of the alzheimer disease in each group). We have then used the knowledge acquired during the first and second part of the assignment towards our main objective set out previously: build a meta-model whose goal is to predict the bias expected when inputting a dataset with a given gender split and some fairness metrics. Each part of the project will here be explained in more detail.

b) First part : Data Investigation with focus against normative outcomes

In developing and pursuing such an ambitious project, we have decided to take an informative and methodical approach to our design. The aim of the first part of the project was to do a statistical analysis of the distribution of Alzheimer disease across a series of features in our dataset.

We first note that the Alzheimer's disease dataset [1] has 373 total entries with an overall gender distribution of 57.1% female and 42.9% male, as seen in Figure 2.

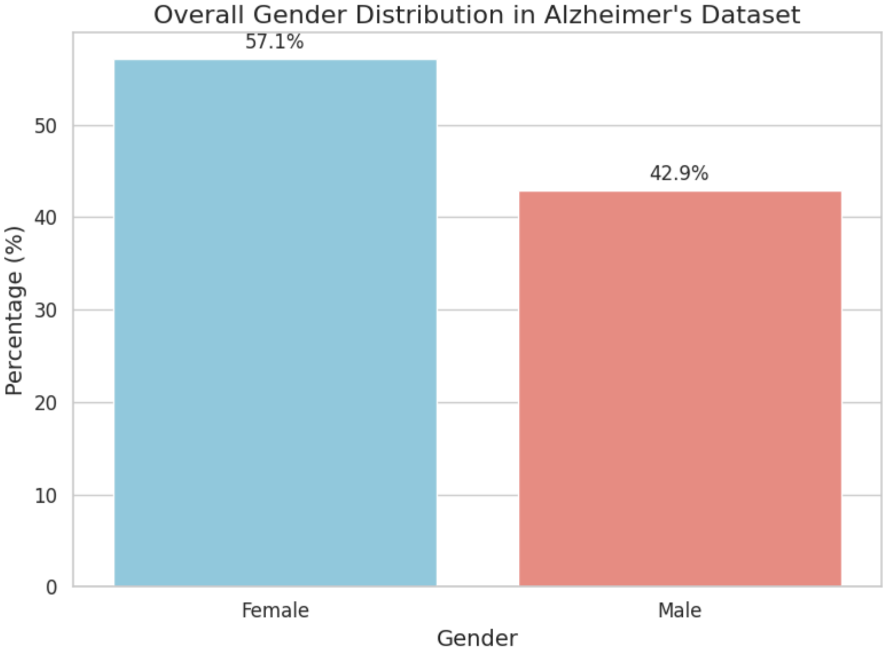
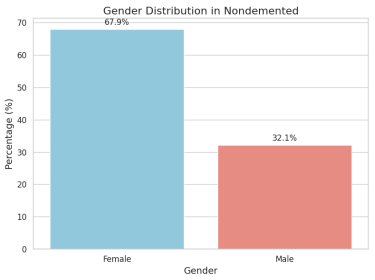
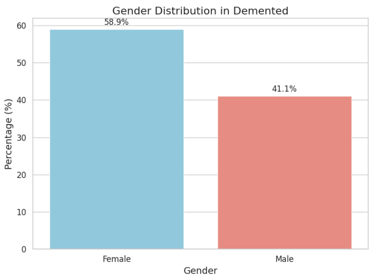
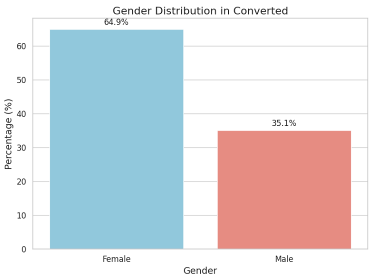


Figure 2: Overall Gender Distribution in Alzheimer's Dataset

It is also important to note that the dataset seems to be a variant of the Oasis dataset [8], yet the exact study from which the data has been extracted is not mentioned anywhere. The participants are classified into nondemented (51%), demented (39%) and converted (10%). From further analysis we can see the gender distribution per group in Figures 3, 4, 5 below.

Figure 3: Gender Distribution in Non-Demented	Figure 4: Gender Distribution in Demented	Figure 5: Gender Distribution in Converted
		

From the figures above, our dataset follows the literary research on which we base our hypothesis. The majority of patients who are demented and convert from nondemented to demented are female. Another critical aspect of the disease is the impact of age and how gender is present within certain age distributions. Therefore, we decided to arbitrarily create age bands of 10 years and plot the count of each group, resulting in Figure 6.

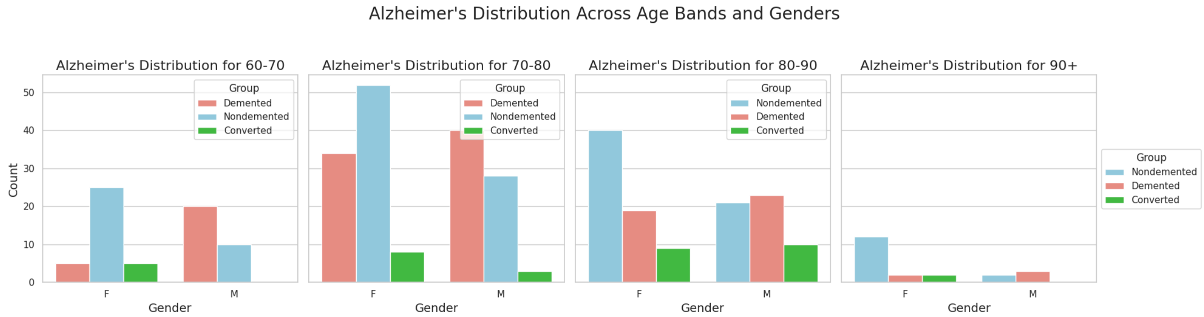


Figure 6: Alzheimer’s Distribution Across Age Bands and Genders

As shown in the figure above, for each age band, you can see the number of individuals in each of the Alzheimer’s disease groups for both genders. This finding is counterintuitive from our previous literary review as more males in the dataset have been classified as demented and thus, such discrepancy can be due to numerous factors such as the origins of the dataset or internal biases of the data that the author has not mentioned.

We decided to investigate these biases further and search whether there were missing entries, which showed us that the dataset is indeed complete, as well as, a chi-square test on each age band to ensure that the distribution observed is statistically significant or due to some random variance in the dataset. Table 1, shows the results of this statistical set on the age bands.

Age Band	Chi-Square Statistic	P-value
60 - 70	20.16	0.000042
70 - 80	6.89	0.031953
80 - 90	4.81	0.090339
90+	4.94	0.084797

Table 1: Chi-Square Statistic on each Age Band

From the table we can deduce the following. The age bands “60 - 70” and “70 - 80” have p-values less than 0.05, suggesting a statistically significant association between gender and Alzheimer’s status. The age bands “80 - 90” and “90+” have larger p-values, so we cannot assert a statistically significant association between gender and Alzheimer’s status for these groups.

Furthermore, we implemented an initial attempt to understand the implication of gender percentage variance in classifiers on the dataset. The target variable was “Group” originally classified, and we deployed Logistic Regression, Decision Trees, and RandomForest classifiers at the current stage. Our metrics for this initial exploratory approach were Accuracy, Precision, Recall, and F1-Score. The results of this analysis are available below:

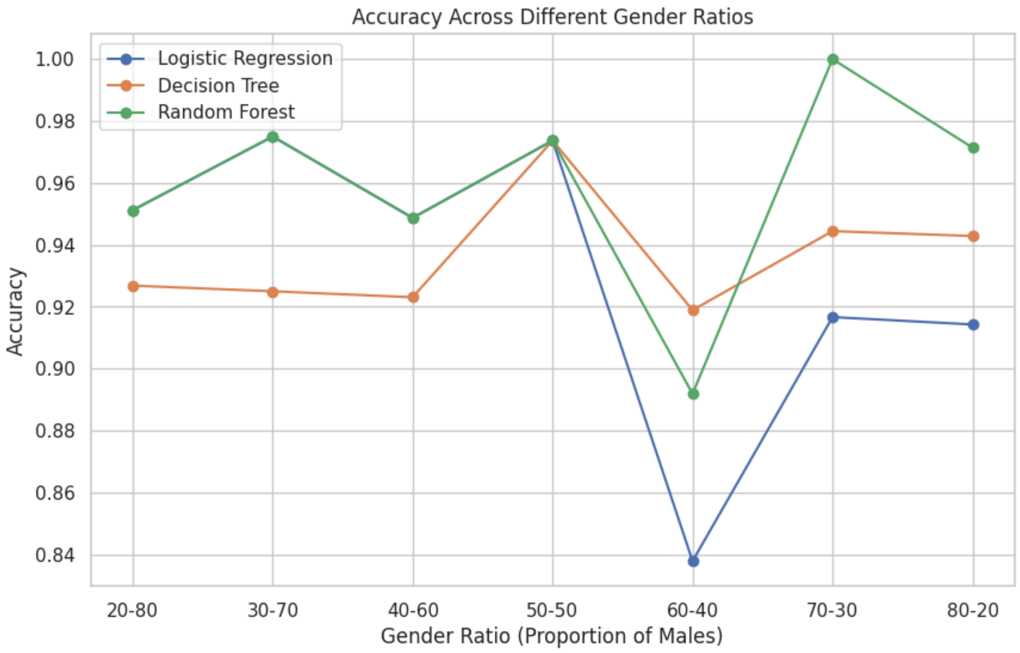


Figure 7: Accuracy Across Different Gender Ratios

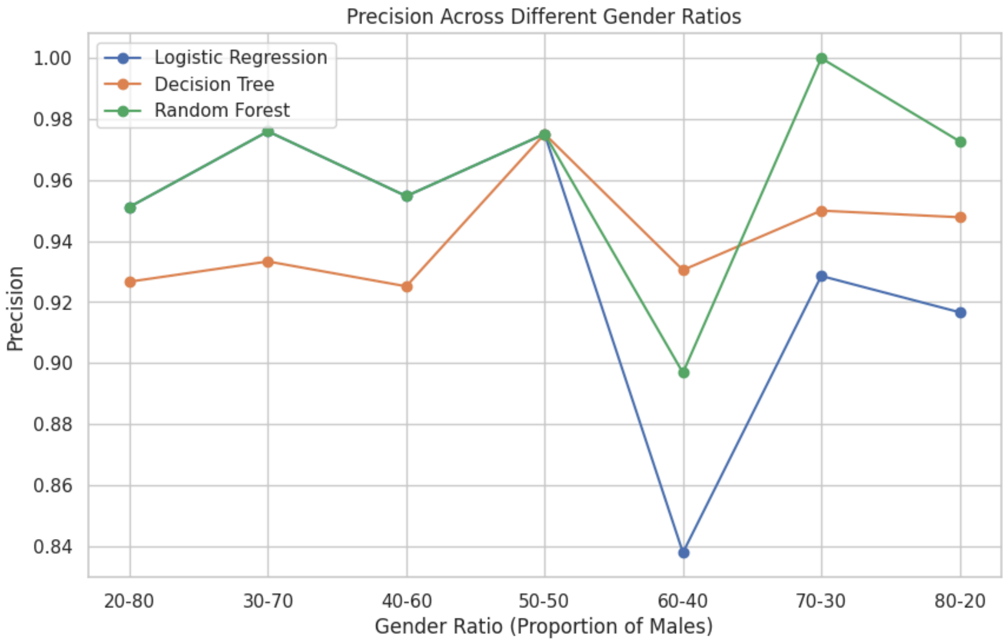


Figure 8: Precision Across Different Gender Ratios

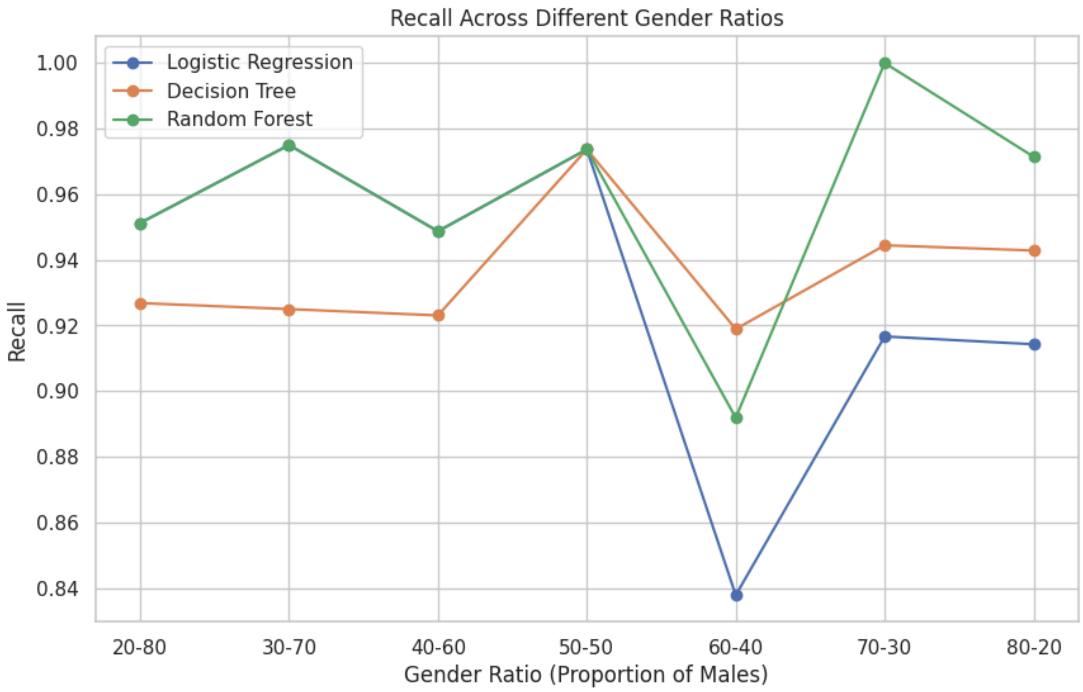


Figure 9: Recall Across Different Gender Ratios

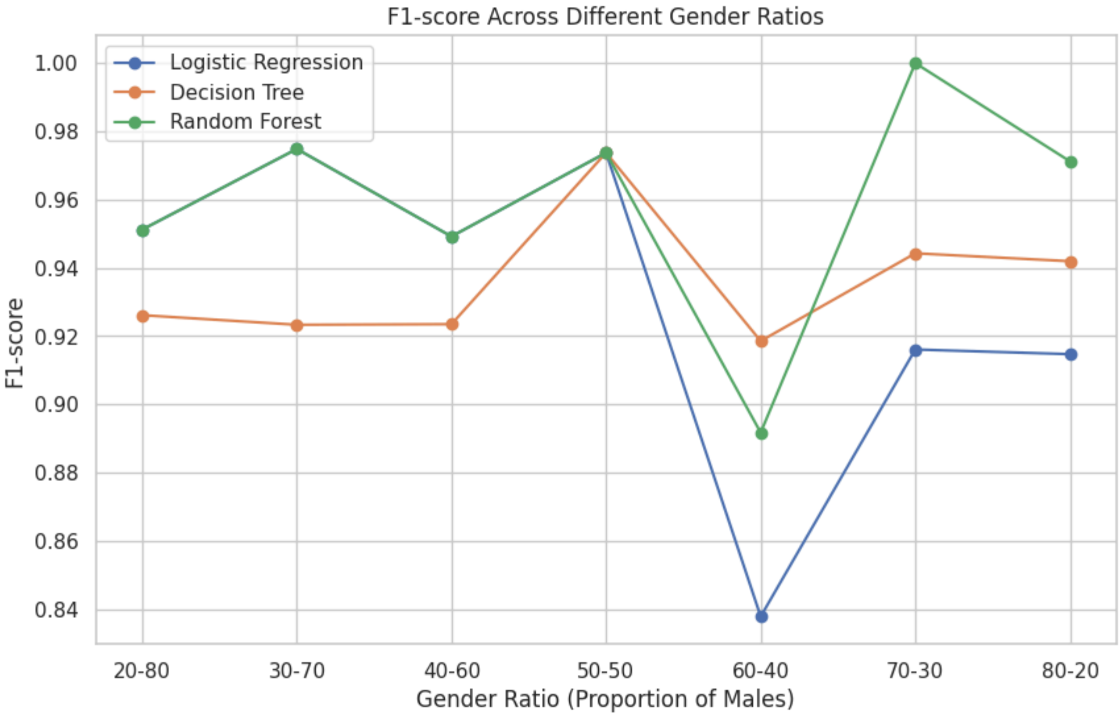


Figure 10: F1-Score Across Different Gender Ratios

From these results, we can conclude that for every metric, there is a noticeable dip for the 60-40 ratio, the majority of men, without any distinct patterns in different distributions. At the 60-40 mark, there could be some overfitting/underfitting by the model which needs to be further explored. The Random Forest seems to be the most robust model showing good performance across all gender ratios. The logistic regression classifier is the most stable and consistent in relation to the others. We also note that all three models achieve the same performance for the 50-50 split.

As opposed to a purely performance-based analysis of the outcomes, our analysis was more focused on the ethical assessment of our design. Therefore, for consistency purposes, we have decided to stack our three models into one stacked classifier, whose end estimator is a logistic regression combining the output of all three models: Logistic Regression, Decision Trees, and RandomForest. Similar accuracy, precision, recall and F1-score patterns were achieved for the stacked model. This way, our analysis was mainly focused on highlighting the effects of varying gender splits on bias rather than analysing only the outcomes of each model. The results of this analysis for the stacked dataset are available below:

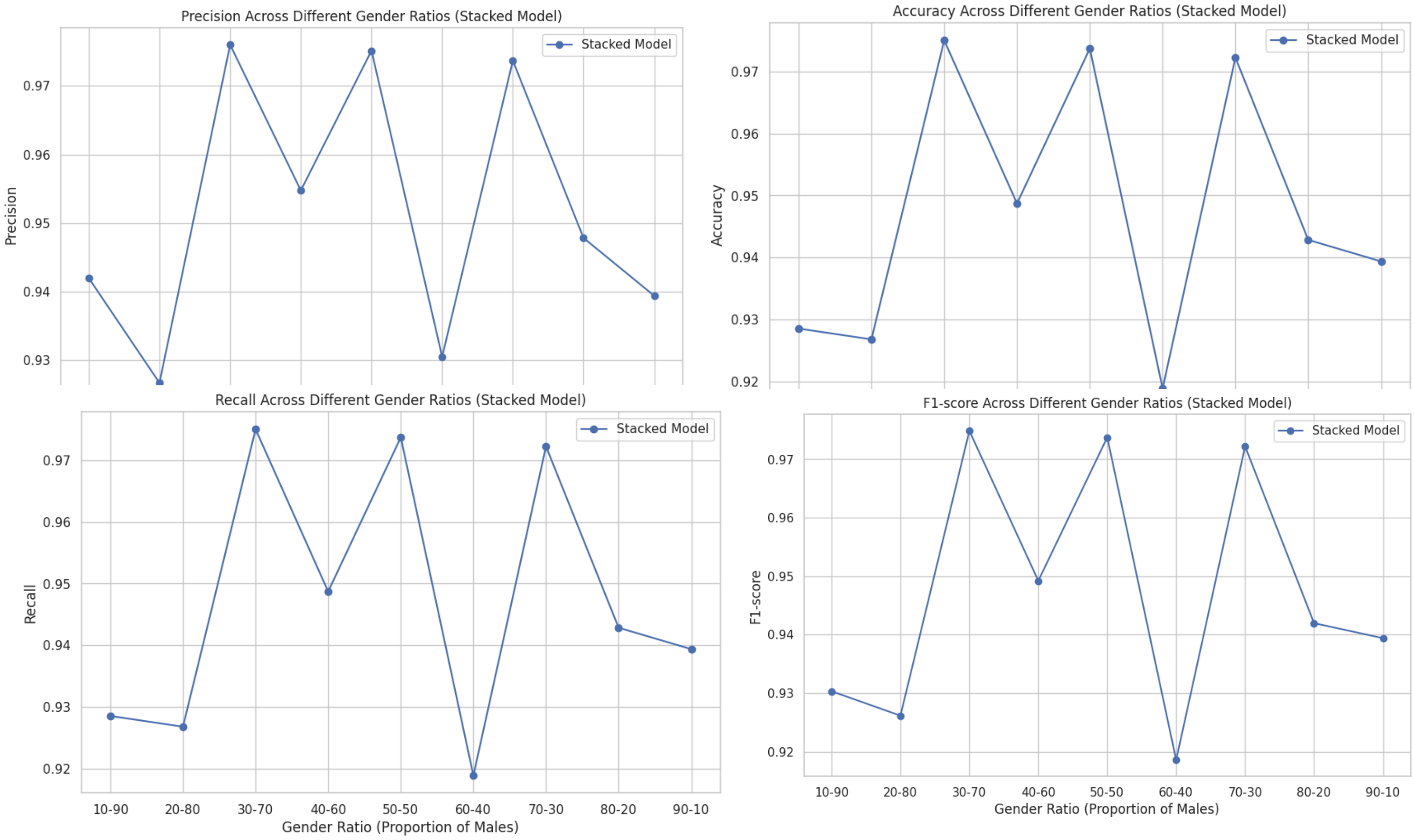


Figure 11: Precision, Accuracy, Recall and F1-score for the stacked classifier

Another investigation we have conducted was building the ROC curves for the stack models trained on each gender split of our dataset. We have also calculated a parameter that we have named Beta, which provides a measure of the area under the curve of the ROC curve for each gender split. The results of this analysis are given below:

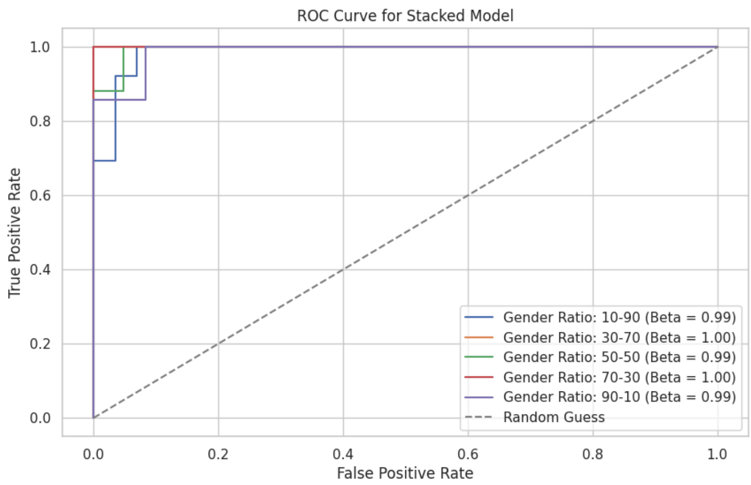


Figure 12 : ROC curves for the stacked models of each gender split

We have also attempted to calculate the bias achieved by each model. First, bias was calculated as the discrepancy between the actual observed value and the anticipated value. We have then plotted the mean squared error (MSE), variance, and bias as calculated above over all gender splits and obtained the following :

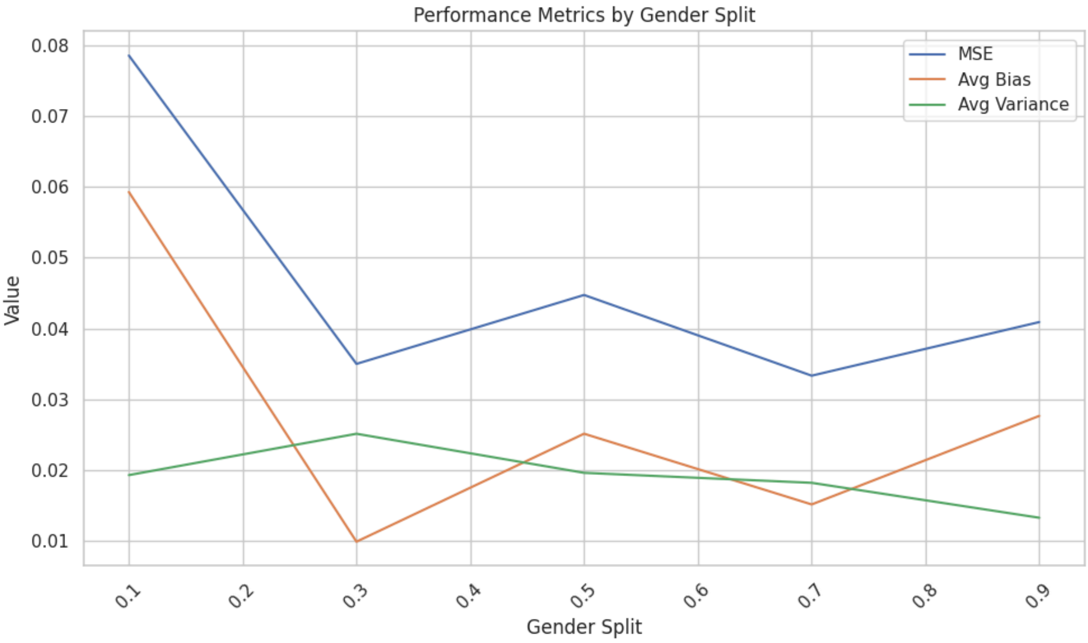


Figure 13 : MSE, variance, and bias across each gender split

Then, because we were more concerned about the impact of varying the gender split on the performance of the model than we are about the absolute discrepancy of output for each model we have built, we have proceeded to change the value of bias that we were calculating. Indeed, the term bias in the following sections of this report will not be used to quantify the discrepancy in output between the predictions of the model and the real labels of the data; it will be used to describe how far away a given performance metric (F1-score, Recall, Precision, Accuracy) achieved by our model is from the average of the metrics of all models.

For instance, if the model developed for a gender distribution of 10% male and 90% female (i.e. a gender split of 0.1) achieves some average value for combined F1-score, Recall, Precision, Accuracy, then the bias is defined as the difference between that average value and the average value of all 4 of those metrics across all gender splits.

To do so, for each gender split, we have built a stacked model that would predict the presence or absence of Alzheimer's disease. Then, we computed some fairness metrics from the AIF360 library for each model. Then, we have calculated the bias of this model.

We plotted the bias values across each gender split for the following series of fairness metrics. We have then observed the following :

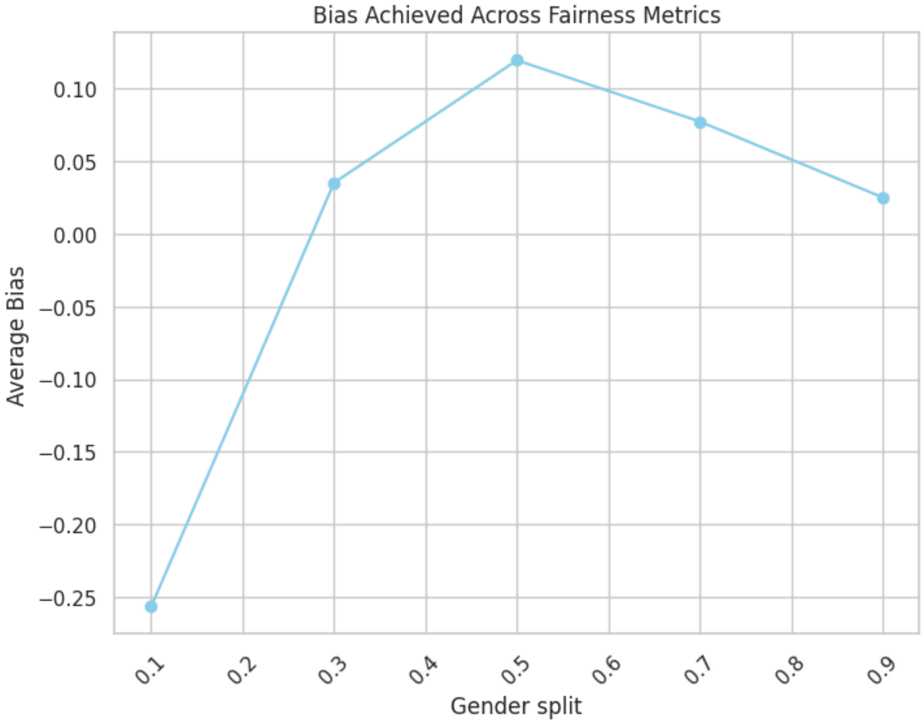


Figure 14 : Bias Achieved For Given List of Fairness Metrics Across Each Gender Split

We have also attempted to reweight each model for completeness purposes using the reweighting preprocessing technique from the AIF 360 library. We have observed some similar profiles of the bias obtained for each gender split, but no significant reduction in the bias was achieved.

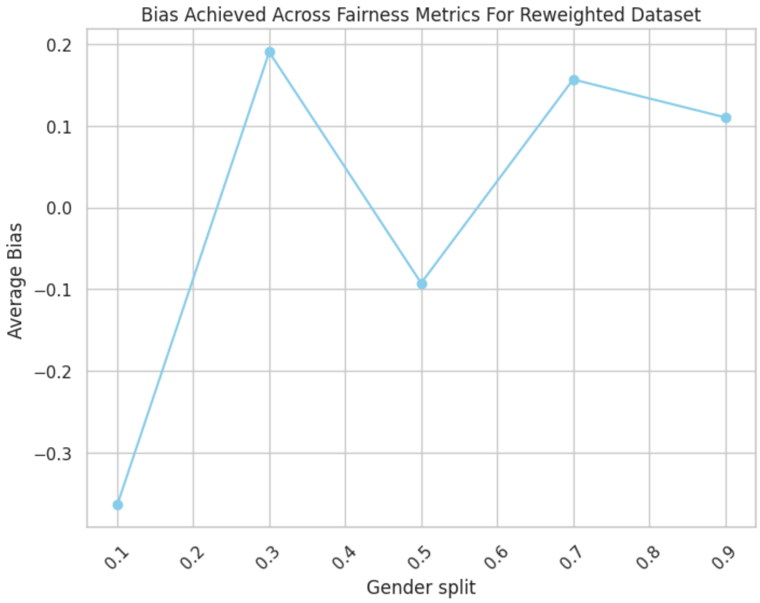


Figure 14 : Bias Achieved For Given List of Fairness Metrics Across Each Gender Split for Reweighted dataset

A more complete breakdown of all the metrics computed across all gender splits is available in the appendix section of this report.

c) Second part towards the design objectives: feature analysis for each group

Our goal in the second section of this report was to highlight which features are the most relevant in predicting Alzheimer's for each group to understand how varying the gender split may affect the model's bias. We conducted a detailed analysis to identify statistically significant features for predicting Alzheimer's disease within each gender group. After segmenting the dataset into male and female subsets, we trained a RandomForestClassifier model on each subset and extracted feature importances. These importances were then visualized using horizontal bar charts, offering insights into the relative importance of different features for Alzheimer's prediction in both male and female populations.

We have achieved the following results :

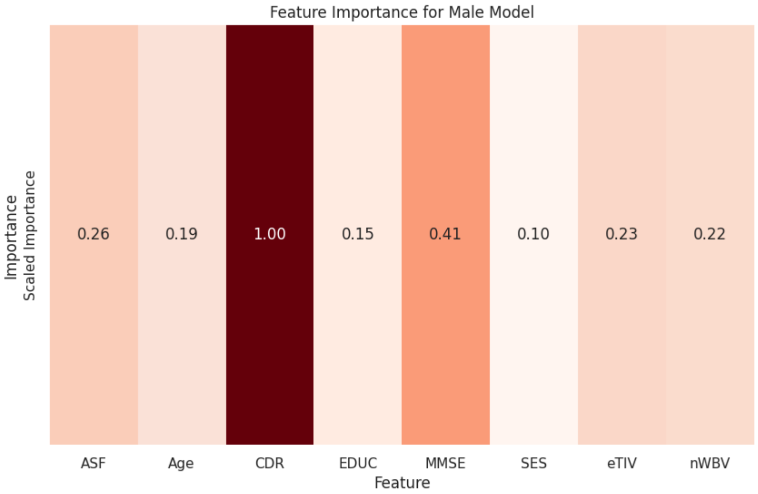


Figure 15 : Feature importance for the Male model

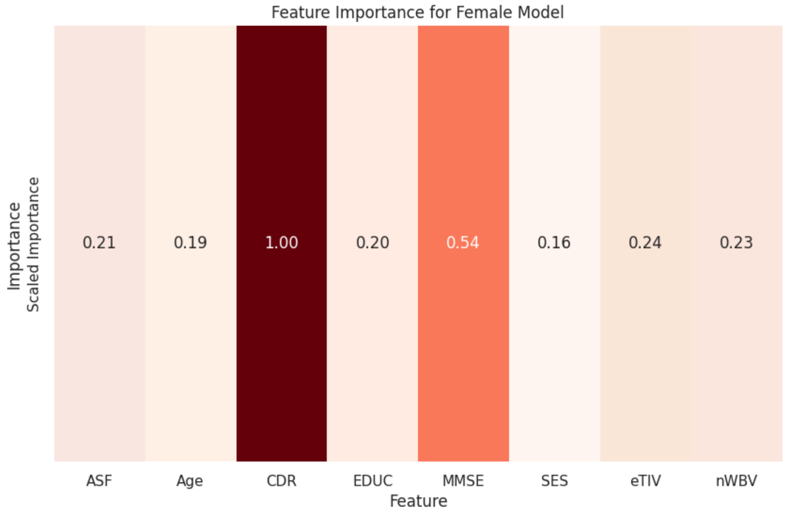


Figure 16 : Feature importance for the Female model

We observe that the CDR feature is the most statistically relevant feature for both groups, and that the second most statistically relevant feature for both groups is MMSE. However, we also observe that this feature is more statistically relevant for the female model than the male model.

d) <u>Third part and final design objective : predicting bias given gender split.</u>

After our in depth statistical analysis of the dataset, we have decided to build a meta-model. Essentially, this model was built to predict the amount of bias to be expected for a given gender split and some fairness metrics describing the dataset onto which our model was trained.

To build this model we have proceeded in the following manner. First, based on the knowledge developed in sections a) and b), we have created a dataset where the features are the values of the fairness metrics achieved for a given gender split and the value of the gender split. It is also relevant to note that we have repeated the previous process for multiple random samples of our dataset : from 20% of the dataset to 80% of the dataset with 10% increments. It is also relevant to note that the dataset we have built is based on the metrics/gender splits of both the Alzheimer and Stroke datasets. This was done to ensure generalisable results. Then, we added the values of bias calculated for each of the gender splits.

Finally, we have built a linear regression algorithm to be able to predict the trends of the expected bias against the real bias for our model. More on this will be discussed in the discussion section of this report. The final results obtained are the following:

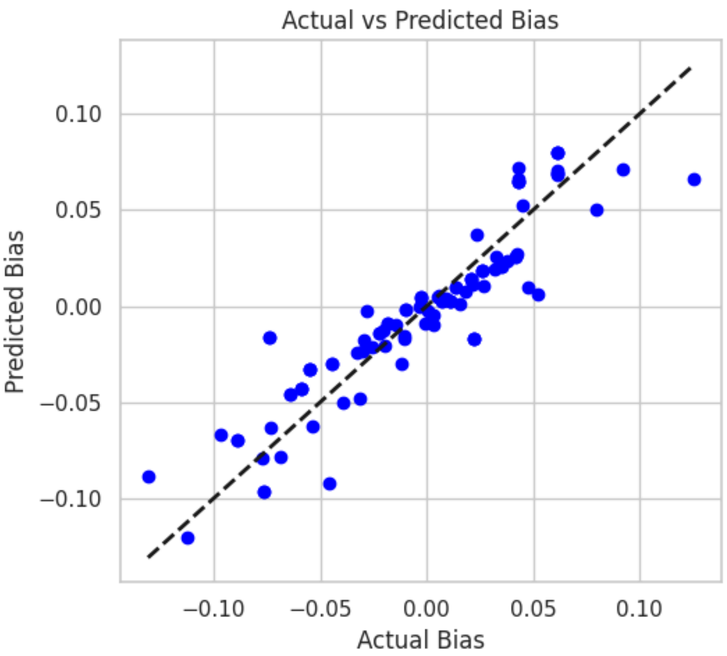


Figure 17 : Regression of Actual vs Predicted Bias by our meta model

We observe that our meta-model is successful in predicting bias when fed some performance metrics and a given gender split, with a given Mean Square error of 0.00039124.

Discussion/Conclusion

This section will focus on the interpretation of the results we obtained by investigations conducted in the methods section.

a) Data Investigation with Focus Against Normative Outcomes:

The initial data analysis of the Alzheimer's dataset revealed interesting insight regarding the distribution of the disease on different demographic groups. Notably, the distribution within the dataset displayed a predominance of females across, which was expected and in-line with existing literature.However, an analysis of age bands highlighted unexpected discrepancies, specifically in the case of higher age brackets, where more males were classified as having Alzheimer's. After further analysis, such as chi-square analysis, more evidence of significant relations between gender and having Alzheimer's within certain age groups. The findings were in-line with our expectations based on our review of existing literature, and highlight the importance of considering demographic factors when interpreting disease prevalence in societal groups.

Moreover, the evaluation of different classifier performance across different gender ratios also resulted in interesting insights. While Decision Tree classifier performed the best in 50/50 gender ratio, in terms of different metrics used for the evaluation of the models (accuracy, F1-score, recall etc.), we could not reach a similar conclusion for the other models including RandomForest, Logistic Regression, and the Stacked Model. We can also note that all models achieve the exact same performance (with regards to all metrics) at a 50-50 gender split. The usage of the Stacked Classifier did not seem to affect the performance of the model across different distributions, apart from allowing for more readable and consistent results.

b) Feature Analysis for Each Group:

A detailed analysis was conducted regarding feature importance for models trained only male patient data and only female patient data. The consistent significance of CDR (Clinical Dementia Rating) and MMSE (Mini-Mental State Examination) underscore their importance in disease prediction regardless of the patient genders. Furthermore, this significance of CDR and MMSE in Alzheimer's diagnosis was also demonstrated previously in certain medical studies [9, 10].

Other features such as ASF (Atlas Scaling Factor), SES (Socioeconomic Status), and EDUC (Education) also demonstrated somewhat significant variations in feature importance between these two models. However, for the case of SES and EDUC combined specifically, the previous studies on the impact of these two factors did not demonstrate significant increased risk between genders or any correlation between adult socioeconomic status [11]. Hence, this variation between feature importance of SES and EDUC between fully-male and full-female trained models can be attributed to algorithmic reasons rather than previously demonstrated underlying medical conditions. However if we had to explain a possible root cause for this reason, as demonstrated by the study in [11], it is possible that female data points had a more prominent presence of patients that had a lower socioeconomic in their early lives which was seen to be more relevant in [11]. In all cases, further research on this topic may be conducted to ensure that the pattern discovered is only localized, and due only to a non-generalized sampling of patients in the dataset.

On the other hand, ASF is a volume-scaling factor associated with equalization of head size [12]. Because of its nature, the impact of ASF is not significantly dependent on gender but rather the physical attributes of a patient. This difference in feature importance can be explained by varying physical attributes between the male and female patients in the dataset and are affected by factors outside the scope of this project. Hence, any relationship between the impact of ASF-Gender can not be drawn from an algorithmic standpoint.

c) Predicting Bias Given Gender Split:

The developed meta-model for bias prediction has been quite successful, as previously mentioned, with a mean square error of 0.00039124 which indicates a high degree of accuracy in the predictions made by the meta-model. By utilizing fairness metrics and gender splits, the meta-model facilitates proactive identification and mitigation of bias, thereby improving the fairness and reliability of predictive models even before they are fully built and trained. Moreover, the incorporation of multiple datasets demonstrates the generalizability of the meta-model across different contexts as well as potential widespread applicability to different diseases.

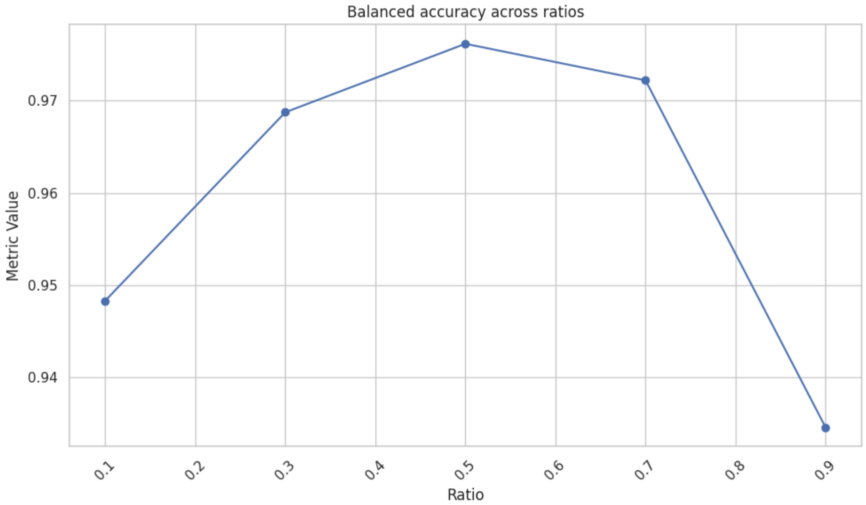
Limitations / Future work:

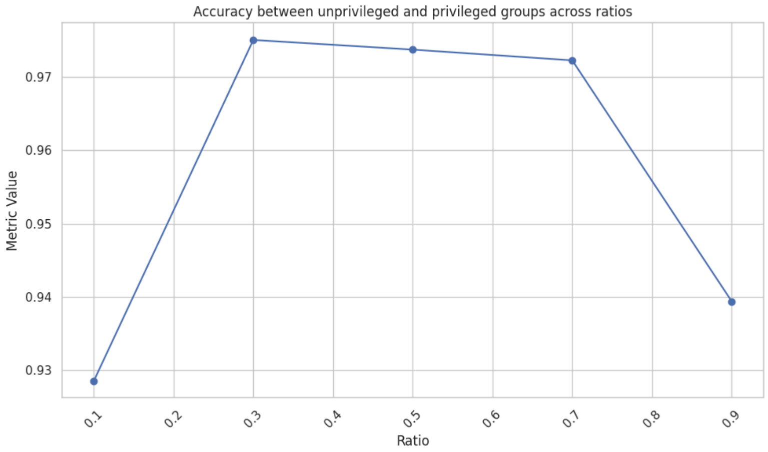
Although the results we obtained in our project are promising there are still areas that can be improved in future work. Firstly, the scalability and adaptability of the meta-model to accommodate additional fairness metrics and datasets requires further exploration. Moreover, the robustness of the meta model against various types of biases beyond gender bias should be investigated as in the context of healthcare and medical applications there are other factors that can have significant potential contributions to bias.Continuous monitoring and updating of the meta-model are also essential to address evolving patterns of bias in real-world datasets.

Another aspect that should be explored in the future is the ethical considerations regarding the use of predicting bias through a meta-model. Evaluation and validation of the meta-model against ethical standards and regulatory guidelines is imperative to prevent unintended consequences and uphold societal trust in ML.

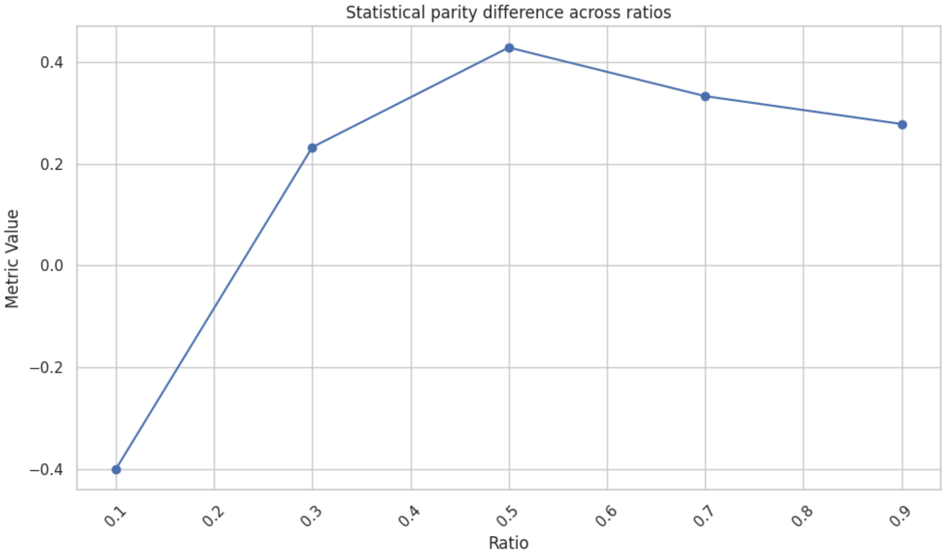
Appendix

Breakdown of all the metrics computed across all gender splits of this report:

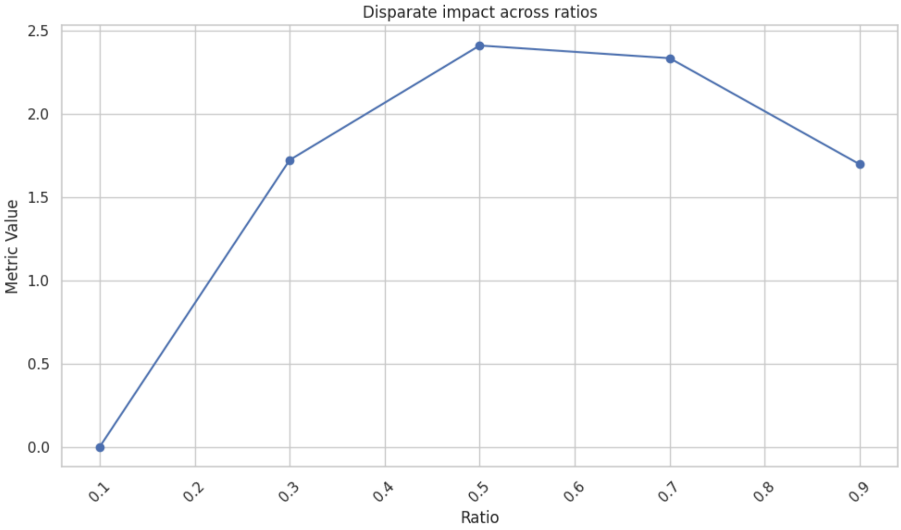




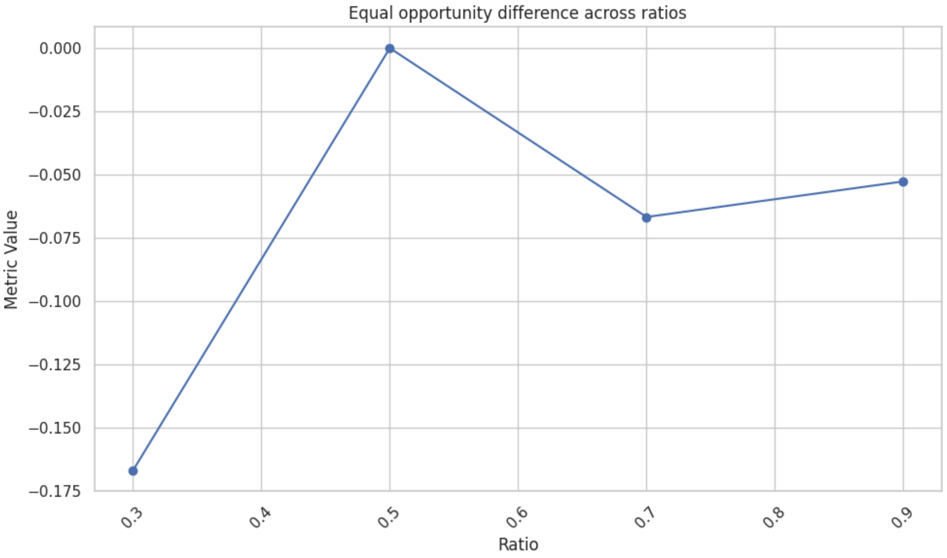
Accuracy between unprivileged and privileged groups across ratios
Metric Value
0.97
0.96
0.95
0.94
0.93
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Ratio



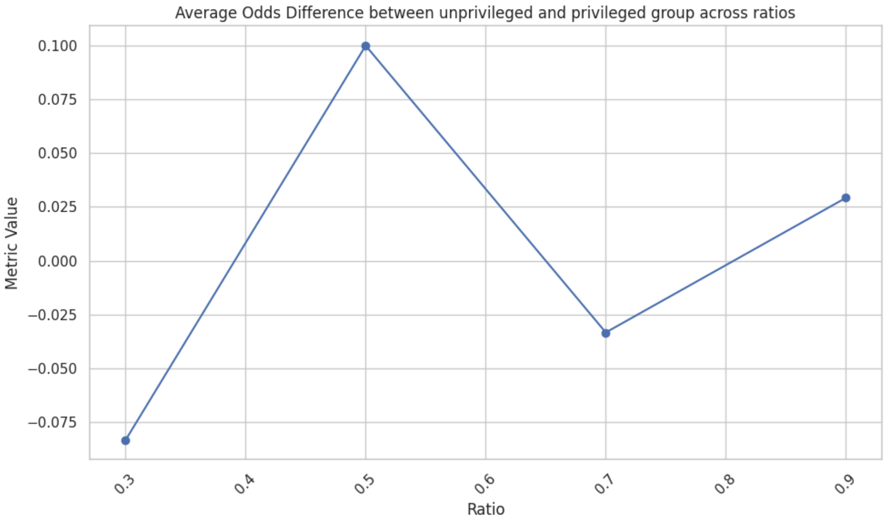
Statistical parity difference across ratios
Metric Value
0.4
0.2
0.0
−0.2
−0.4
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Ratio



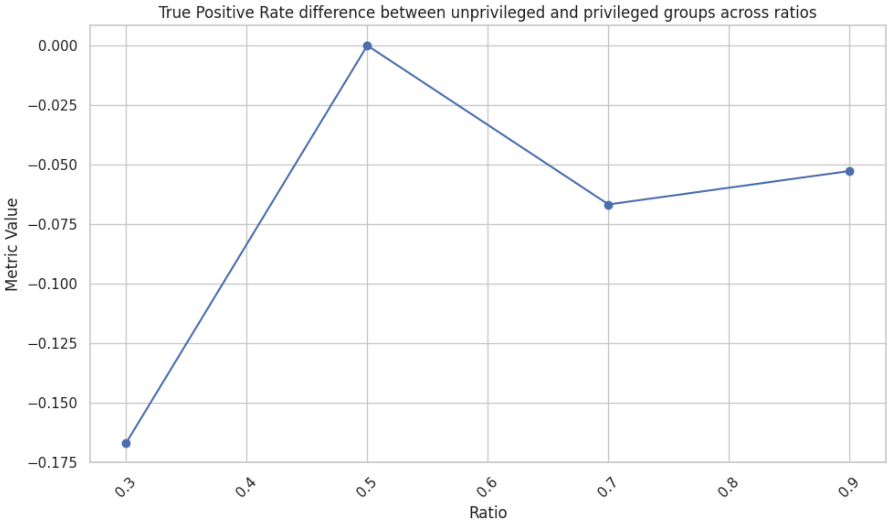
Disparate impact across ratios
Metric Value
2.5
2.0
1.5
1.0
0.5
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Ratio



Equal opportunity difference across ratios



Average Odds Difference between unprivileged and privileged group across ratios



True Positive Rate difference between unprivileged and privileged groups across ratios

References

[1] “Alzheimer Features,” *www.kaggle.com*. https://www.kaggle.com/datasets/brsdincer/alzheimer-features

[2] FEDESORIANO, “Stroke Prediction Dataset,” *www.kaggle.com*, 2021. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[3] “Women and Alzheimer’s,” Alzheimer’s Disease and Dementia, 2024. https://www.alz.org/alzheimers-dementia/what-is-alzheimers/women-and-alzheimer-s#:~:text=Women%20at%20risk-

[4] A.E. Budson, “Why are women more likely to develop Alzheimer’s disease?” Harvard Health Blog, Jan. 2022. Available: https://www.health.harvard.edu/blog/why-are-women-more-likely-to-develop-alzheimers-disease-202201202672

[5] K. M. Rexrode, T. E. Madsen, A. Y. X. Yu, C. Carcel, J. H. Lichtman, and E. C. Miller, “The Impact of Sex and Gender on Stroke,” Circulation Research, vol. 130, no. 4, pp. 512–528, Feb. 2022, doi: https://doi.org/10.1161/circresaha.121.319915.

[6] J. S. Suri *et al.*, "Five Strategies for Bias Estimation in Artificial Intelligence-based Hybrid Deep Learning for Acute Respiratory Distress Syndrome COVID-19 Lung Infected Patients using AP(ai)Bias 2.0: A Systematic Review," in *IEEE Transactions on Instrumentation and Measurement*, doi: https://doi.org/10.1109/TIM.2022.3174270

[7] Chung H, Park C, Kang WS, Lee J. Gender Bias in Artificial Intelligence: Severity Prediction at an Early Stage of COVID-19. *Front Physiol*. 2021;12:778720. Published 2021 Nov 29. doi: https://doi.org/10.3389/fphys.2021.778720

[8] Kaur, H., Pannu, H.S., & Malhi, A. K. (2019) A systematic review on imbalanced data challenges in machine learning: Applications and solutions. ACM Comput. Surv., 52(4), 1-36.

[9] Perneczky, R., Wagenpfeil, S., Komossa, K., Grimmer, T., Diehl, J., & Kurz, A. (2006). Mapping scores onto stages: mini-mental state examination and clinical dementia rating. The American journal of geriatric psychiatry : official journal of the American Association for Geriatric Psychiatry, 14(2), 139–144.

[10] Chapman, K.R., Bing-Canar, H., Alosco, M.L. et al. (2016). Mini Mental State Examination and Logical Memory scores for entry into Alzheimer's disease trials. Alz Res Therapy 8, 9.

[11] Karp, A., Kåreholt, I., Qiu, C., Bellander, T., Winblad, B., & Fratiglioni, L. (2004). Relation of education and occupation-based socioeconomic status to incident Alzheimer's disease. American journal of epidemiology, 159(2), 175–183.

[12] Buckner, R. L., Head, D., Parker, J., Fotenos, A. F., Marcus, D., Morris, J. C., & Snyder, A. Z. (2004). A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume. NeuroImage, 23(2), 724–738.

[13] Nataly Buslón, Àtia Cortés, Silvina Catuara-Solarz, D. Cirillo, and María José Rementería, "Raising awareness of sex and gender bias in artificial intelligence and health," Frontiers in global women's health, vol. 4, Sep. 2023, doi: https://doi.org/10.3389/fgwh.2023.970312.

[14] D. Cirillo et al., "Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare," npj Digital Medicine, vol. 3, no. 1, Jun. 2020, doi: https://doi.org/10.1038/s41746-020-0288-5.

[15] D. Cirillo and M. J. Rementeria, "Chapter 3 - Bias and fairness in machine learning and artificial intelligence," ScienceDirect, Jan. 01, 2022. https://www.sciencedirect.com/science/article/abs/pii/B9780128213926000066 (accessed Mar. 14, 2024).

[16] S. Raza, "A machine learning model for predicting, diagnosing, and mitigating health disparities in hospital readmission," Healthcare Analytics, p. 100100, Aug. 2022, doi: https://doi.org/10.1016/j.health.2022.100100.

[17] K. Hariharan, "How Will AI Affect Gender Gaps in Health Care?," www.mmc.com. https://www.marshmclennan.com/insights/publications/2020/apr/how-will-ai-affect-gender-gaps-in-health-care-.html

[18] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," ACM Computing Surveys, vol. 54, no. 6, pp. 1–35, Jul. 2021, doi: https://doi.org/10.1145/3457607.

[19] "OASIS: Online Atlas of Shared Interests and Structures," Online: http://www.oasis-brains.org/

[20] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

[21] Elkan, C. (2001). The foundations of cost-sensitive learning. In Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2 (IJCAI'01).

[22] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.