Introduction

Artificial-intelligence (AI) integration into healthcare is an emerging field of innovation that has the potential to greatly improve patient care and expedite medical diagnosis and treatment procedures (Pesapane, Codari and Sardanelli, 2018). This technical progress is not without its difficulties, though. The inherent bias in AI-systems is one of the biggest issues since it can have a big impact on how diseases like heart-disease are diagnosed and treated. AI-bias is a serious challenge to the objective and equitable application of AI in healthcare (Topol, 2019). It can originate from a variety of sources, including data-skewness, algorithmic prejudice, and subjective human input during the training-process.

Inconsistencies in the outcomes of medical treatments can be strengthened and maintained by biased AI-models. For instance, an AI-system may function less well for people outside of a given demographic group if it was trained mostly on data from that group. With heartdisease, early identification and precise risk-factor evaluation are essential for successful treatment and management, thus this is especially concerning. The lack of diversity in ethnic backgrounds, gender, and socioeconomic level in the data might lead to biased predictions being made by AI-systems, as (Kent Baker et al., 2020) showed. These biases may result in underprivileged populations who are already disadvantaged in terms of healthcare outcomes and access receiving subpar care. (Agarwal et al., 2022) examined a noteworthy example that brought to light racial bias in an algorithm that is used for healthcare-management nationwide. Due to skewed training-data reflecting historical disparities in healthcare access and utilization, the study found that the algorithm consistently overestimated the health needs of Black-patients relative to White-patients.

Also, it is impossible to overestimate the moral ramifications of biased AI decision-making in the healthcare industry. Since AI is used to make crucial health choices, any underlying prejudice raises concerns about both the equity of healthcare services provided and the effectiveness of AI-systems. To guarantee that all patients, regardless of background, can fairly and equally benefit from AI-technology, addressing these biases is both a technical problem and a moral obligation.

This research will examine the causes and effects of bias in AI-driven models for heartdisease to address these problems. It will look at how inclusive data-collecting, algorithmic transparency, and stringent validation procedures might foster justice in AI. Healthcareproviders and AI-developers can collaborate to use AI's potential in a way that advances health equity and enhances patient outcomes by putting these tactics into practice.

Data-Preprocessing

Dataset-Overview: The dataset is substantial, with 445,132 entries before cleaning and 39 different variables, ranging from demographic data to detailed health-indicators. The source of the dataset was originated from Kaggle Indicators of Heart Disease (2022 UPDATE) (kaggle.com)

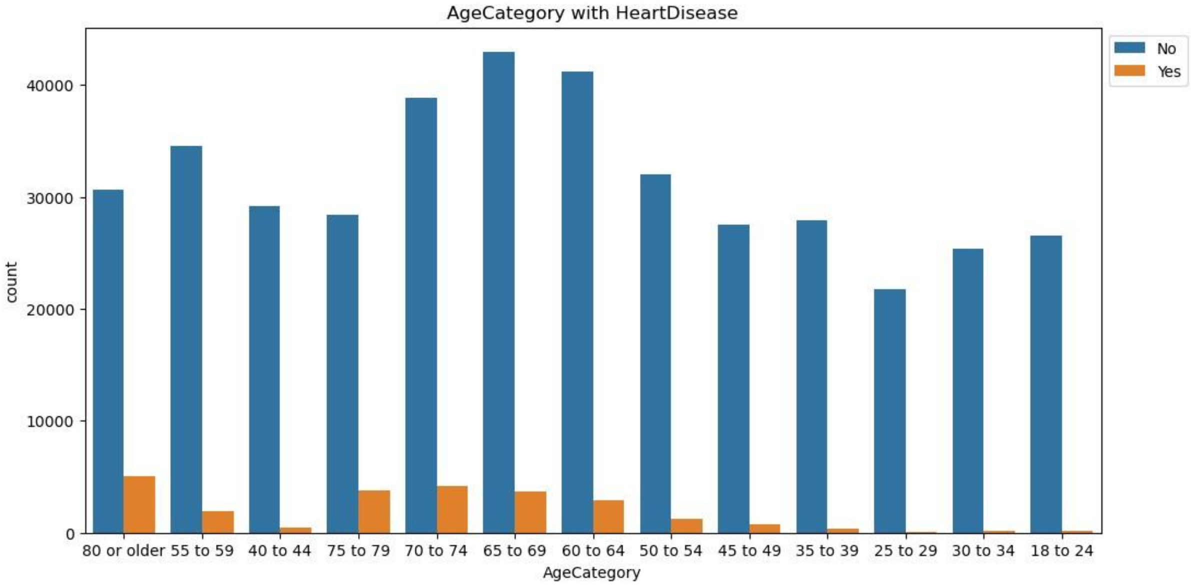
Variable-Overview: The dataset consists of Demographic and Health-Variables, HealthStatus Indicators, Lifestyle-Indicators, and Miscellaneous-Health Data.

Managing-Missing Data: To maintain the central-tendency and structure of the data, missing values in numerical-columns were imputed using each column's median. This is a useful technique for skewed-distributions that are common in clinical-data. In the case of categorical-columns, the distribution integrity of the variables was preserved, and existing data patterns were adhered to by imputed values by using the mode to fill in the missingvalues.

Handling-Duplicate and Measuring-Dimensions: A total of 1,693 duplicates were found in the dataset; the first duplicate is retained, while any further instances are eliminated. The total number of elements in the dataset is represented by the size of the dataset, which was 17,294,121.

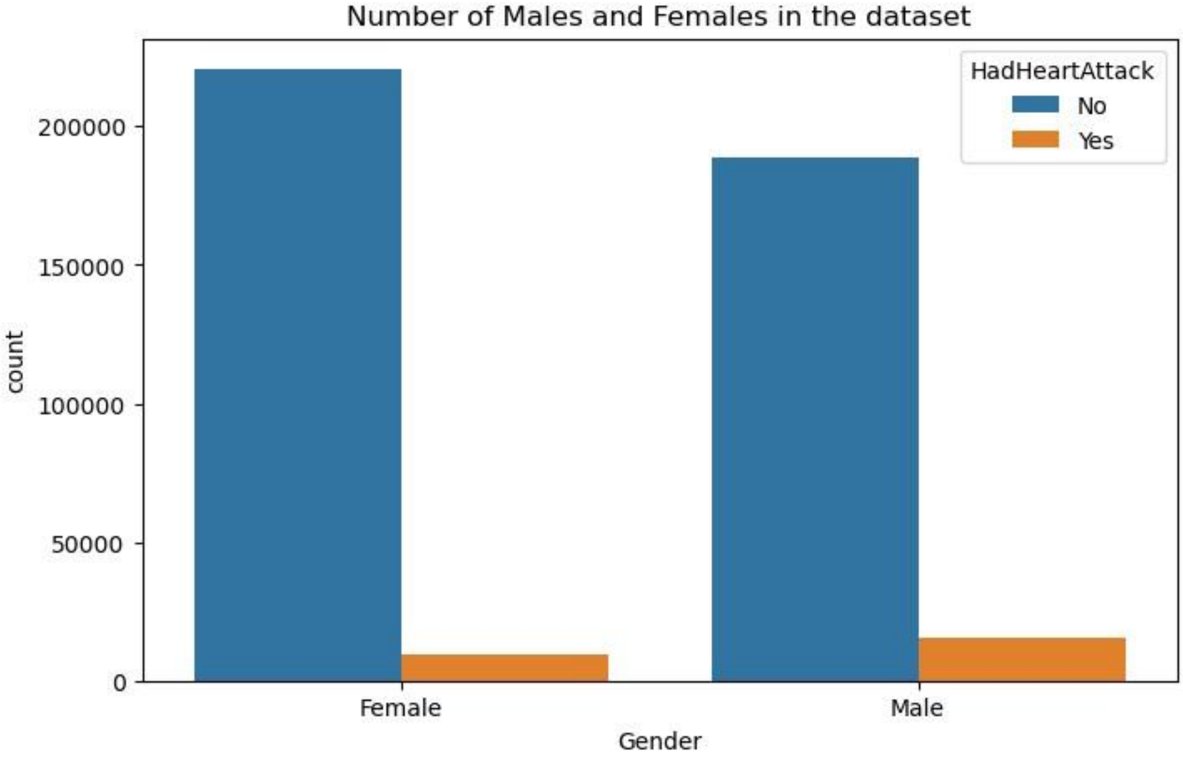
Exploration-Data-Analysis

Age-Category and Heart-Disease Prevalence Increase with Age: The orange-coloured heart-disease instances increase with age, corresponding to the established medical-risk increase associated with aging. The blue-bars indicate that although heart-disease is common, most of the sample consists of people who do not have the illness. The older agegroup shown by a larger orange segment has the highest relative prevalence of heartdisease. Conversely, younger-groups whose orange segments are smaller have noticeably fewer cases. Comprehending these age-related trends is essential for customizing healthstrategies and optimizing prediction algorithms.



Gender-Disparity in Heart-Disease: Men are more likely than women to experience heartattacks, which is a notable gender-disparity in heart-disease incidence. 187,527 men and 219,411 women do not have heart-disease, but 15,198 men and 9,549 women have.

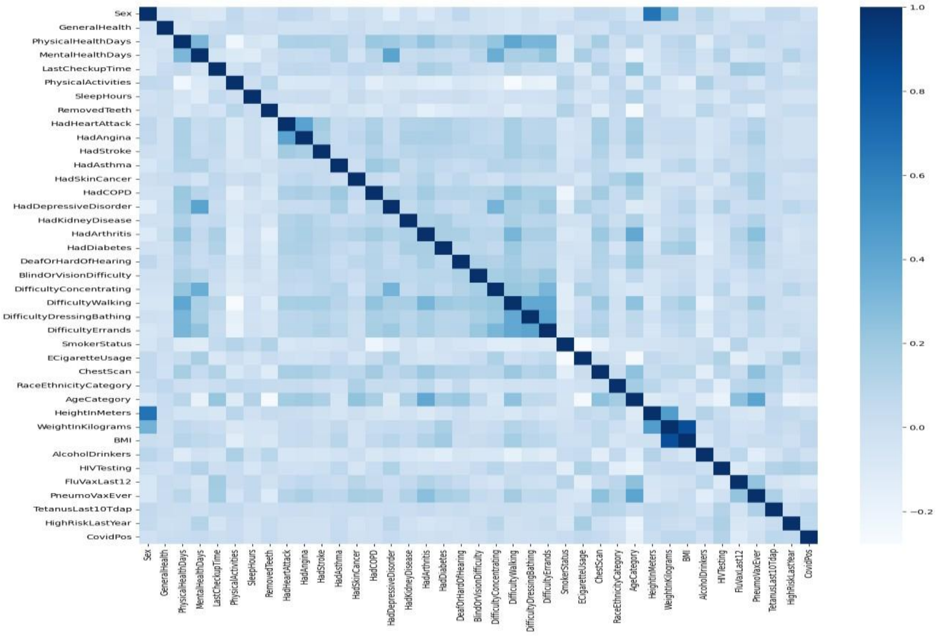
According to study, men are generally more prone to heart-disease across most age-groups, which implies that men are at a higher risk.



Predictive Modeling Implications

Predictive model-development has been informed by the chart findings: It is important to emphasize age, as it is a powerful indication of heart-disease risk. For an accurate riskassessment, gender-differences imply that models may need to be adjusted or stratified. Before training, methods such as SMOTE are crucial to rectify class-imbalance because of the unequal distribution of heart-disease, especially in younger age-groups.

Correlation-Heatmap-Analysis



Strong correlations between closely-related variables such as those between health-related variables like HadHeartAttack, HadAngina, and HadStroke are shown by darker colours in clustering. These correlations are important for predictive models since significant multicollinearity (correlation between Independent-Variables) can reduce precision and statistical power. On the other hand, high correlations between the independent and dependent-variables imply that the former are highly predictive of the latter, such as heart disease.

Feature-Encoding and Scaling: By encoding them in a way that machine-learning algorithms can easily interpret, such as label-encoding, categorical-variables are made more readable. Standardization is also used to scale numerical characteristics, normalizing the data and accelerating the convergence of training algorithms.

Data-Resampling: The model employs both under and oversampling strategies to solve class-imbalance, a common problem in medical datasets where outcomes such as heartdisease are underrepresented. While SMOTE (Synthetic-Minority Over-sampling-Technique) synthesizes additional samples in the minority-class to enhance the model's learning and aim for a balanced dataset (Tang et al., 2023), Random-Under-Sampler minimizes the size of the majority class to prevent bias.

Model-Development

The model is developed using a methodical approach to selecting and assessing machinelearning algorithms:

Choosing-Algorithms: Gradient-Boosting-Classifier, which can model complex relationships and is resistant to outliers (Khera et al., 2021), is used in this work. This method is selected for medical prediction jobs due to its feature handling capability and ability to handle non-linear data.

Pipelines- Integration: All changes stay consistent between training and testing when these models are integrated into a pipeline with preprocessing stages. The integrity of the model's performance depends on this methodological rigor.

Accuracy-scores, classification-reports, and confusion-matrices are used to assess the model's performance. These measures offer a thorough understanding of the model's efficacy, revealing not only the overall accuracy but also the class's precision, recall, positive-rate, and F1-score. ROC-curves are used to evaluate the model's class discrimination ability, and the AUC-score measures the model's overall efficacy. Improved model-performance is shown by higher AUC-values.

Application of Fairness-Criteria

An investigation of fairness focuses on how equally the model's predictions apply to various groups that are defined by protected characteristics, like gender. Preventing discriminatory outcomes in healthcare-settings is crucial. Important facets of analysis of fairness comprise:

Gender-spratesic Performance-Metrics: The model's accuracy, recall, precision, and positive-rate are computed individually for each gender following the prediction of heartdisease outcomes.

Equivalent-Precision

Comparative-Analysis: To determine whether the model works equally well for all groups, accuracy is tested independently for each group (such as males and females).

Discrepancies-in-Accuracy: Any significant difference in accuracy between groups could suggest that the model is biased, favouring one group over another.

Equal-Opportunity (Recall)

This statistic, which assesses whether every group has an equal likelihood of being correctly recognized as positive, is essential for fairness. A poorer recall for a particular group may indicate that the model is less effective in identifying heart-disease within that group. **Gender-Group ROC-Curves**: To visually compare the model's discrimination threshold for each group, separate ROC-curves are presented for men and women. A measurable indicator of the model's gender-neutral predictive parity are the matching AUC-ratings. **Demographic-Parity (Positive-Predictive-Value (PPV) and Positive-Rate**): **PPV**: Determines whether the percentage of real positives among the cases that are anticipated to be positive is consistent across various groupings.

Positive-Rate: Examines whether the rate of accurate predictions varies across various demographic groups to make that the model does not predispose any one group to a higher risk of heart-disease than another.

The Gradient-Boosting model's effectiveness and equity are guaranteed by a thorough review of the model using fairness-analysis and performance measures. This dual approach is especially important in healthcare applications where there are high risks since it aids in spotting potential biases and taking corrective action to mitigate them. It is advised that the model be updated and monitored continuously considering these assessments to preserve its fairness and accuracy throughout time.

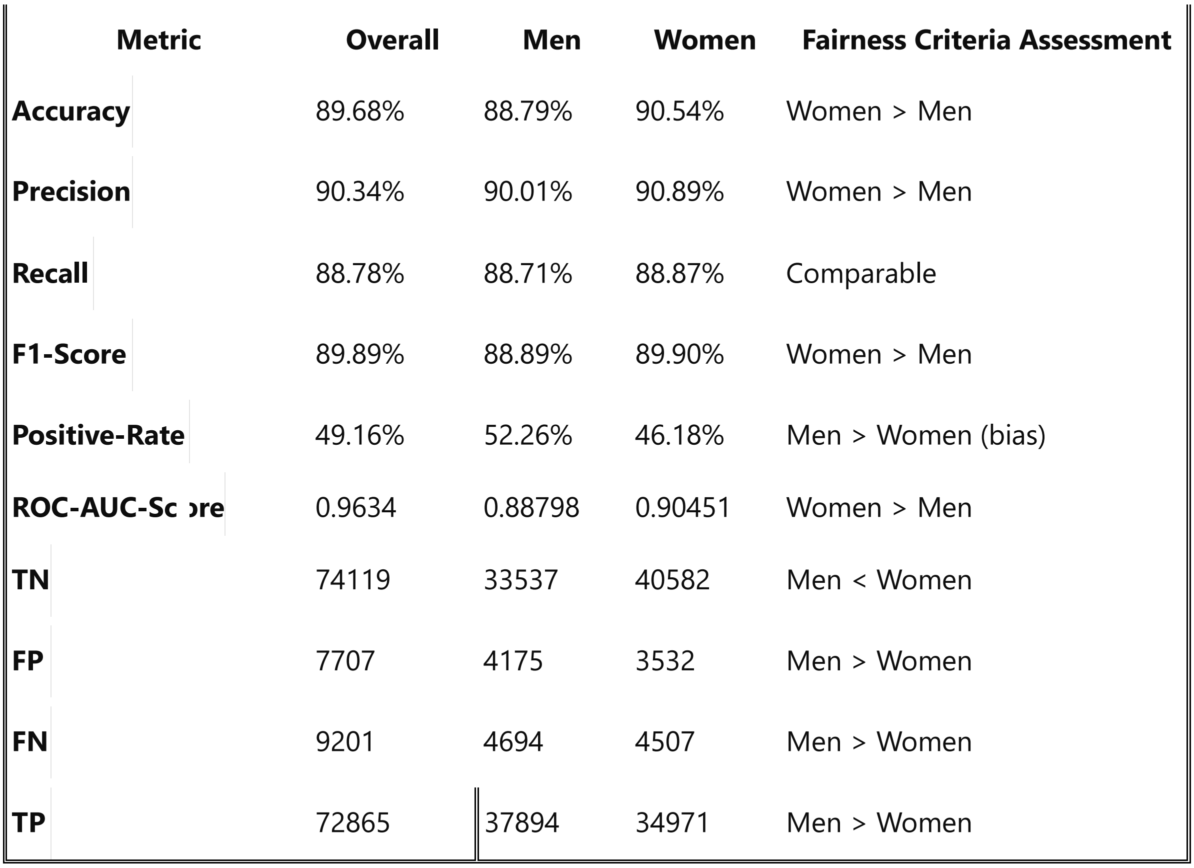
This study evaluates the gender-neutrality of the Gradient-Boosting classifier used to predict heart-disease. Determining if the model shows any bias that might affect the fair treatment of patients requires an understanding of these inequalities.

Result

Fairness-Analysis and Model-Performance

Model-Performance Overall: With accuracy and balanced precision and recall across the predicted classes, the Gradient-Boosting classifier shows great overall efficacy. Excellent model-performance in differentiating patients with and without heart-disease is shown by the ROC-AUC-Score.

Model-Performance and Fairness-Analysis Table



Metric	Overall	Men	Women	Fairness Criteria Assessment
Accuracy	89.68%	88.79%	90.54%	Women > Men
Precision	90.34%	90.01%	90.89%	Women > Men
Recall	88.78%	88.71%	88.87%	Comparable
F1-Score	89.89%	88.89%	89.90%	Women > Men
Positive-Rate	49.16%	52.26%	46.18%	Men > Women (bias)
ROC-AUC-Score	0.9634	0.88798	0.90451	Women > Men
TN	74119	33537	40582	Men < Women
FP	7707	4175	3532	Men > Women
FN	9201	4694	4507	Men > Women
TP	72865	37894	34971	Men > Women

Analysis of Bias

Performance-Disparity: The model shows that women perform with greater accuracy and precision than males do. This discrepancy implies that women's cases of heart-disease are more accurately identified and confirmed by the model than those of males.

Equal-Opportunity (Recall): Men and women had remarkably similar recall rates, indicating that the model's sensitivity in identifying genuine positive cases is equivalent for both genders. In huge datasets, this slight variation can nonetheless have a big impact.

Demographic-Parity (positive-rate): Males have a larger positive-rate than females, this suggesting that men are more likely to have heart-disease to be predicted by the model. Demographic-parity dictates that every group should have an equal chance of obtaining a favourable forecast, which is broken by this.

ROC-AUC-Score: A higher-score for women means that the model is more effective at differentiating between heart-disease and non-heart-disease. This raises the possibility of model calibration problems, in which case the gender-specific threshold for forecasting positive cases may not be appropriately set.

Fairness-Criteria Not Met

From the bias-analysis, it is evident that the model fails to meet several key fairness-criteria:

Demographic-Parity: The disparity in positive-rates between men and women indicates that the model does not meet demographic-parity. All groups should have a similar percentage of positive predictions in a fair model so that no group is unfairly labelled as having the illness.

Equalized-Odds: The model does not fully satisfy equalized-odds, despite the near recallrates. This is indicated by the little discrepancy along with the variations in positive-rates and ROC-AUC-scores. This does not entirely accomplish equalized-odds, which call for equal true-positive and false-positive rates throughout groups.

Predictive-Equality focuses on Equal false positive-rates: Predictive equality may have been violated because of the greater positive-rate for men, which suggests that the false positive-rate is likewise higher for this group.

Clinical and Ethical-Consequences

Male-Overtreatment Risk: Higher false positive rates in men may result in overtreatment, putting patients through pointless tests or treatments that have financial ramifications and put them at unnecessary danger.

Women-Undertreatment Risk: Despite the high ROC-score, the lower positive-rate in women advises caution when predicting heart-disease, raising the possibility of undertreating real cases. Delays in critical actions that are necessary to stop the progression of serious diseases might be caused by undertreatment.

Trust and Reliability: The confidence that patients and healthcare professionals have in AIbased diagnoses may be impacted by variations in model-performance. Doubt over the application of such models in clinical practice can stem from perceived biases or mistakes.

Regulatory and Legal-Concerns: Models that exhibit bias or uneven performance amongst protected categories, such as gender, may come under regulatory scrutiny, which could have an impact on their acceptance and application in clinical settings.

Balanced Decision-Making: When using AI-predictions to make clinical choices, medical professionals must be conscious of these biases. This understanding is essential to coordinating AI inputs with other diagnostic data and patient-specific factors.

Conclusion

The analysis indicates that the model slightly favours female-patients in terms of accuracy, precision, and discriminative ability (ROC-AUC). However, it also shows a tendency to predict heart-disease more frequently in men, which may lead to a higher false-positive-rate. This gender-based disparity underscores the need for model recalibration or the development of gender-specific models to ensure equitable healthcare outcomes.

The higher-performance metrics for women suggest that the model is tuned more effectively to the characteristics and patterns prevalent in the female data it was trained on. Conversely, the higher positive-rate in men raises concerns about potential overdiagnosis or overtreatment, which could have significant implications for patient-care and resource allocation in healthcare settings.

Recommendations

To mitigate these biases and enhance model fairness:

Rebalance the Dataset: Adjust the training-data to ensure balanced representation and treatment of both genders in the learning process.

Feature-Engineering: Reevaluate the input features for gender-specific biases that might influence model decisions.

Algorithmic-Adjustments: Explore alternative algorithms or adjust model parameters to minimize disparities in predictions across genders.

Continuous-Monitoring: Regularly review the model’s performance and impact, particularly looking at long-term health outcomes for all demographics to ensure sustained fairness.

Future-Studies

To improve the model’s fairness and generalizability, further research should include more diverse datasets with a wider variety of clinical and demographic-factors.

More research may be done on advanced modeling strategies and fairness-enhancing algorithms to reduce bias and boost forecast-accuracy across various demographic groupings.

Reference

Agarwal, R., Bjarnadottir, M., Rhue, L., Dugas, M., Crowley, K., Clark, J. and Gao, G. (2022). Addressing Algorithmic Bias and the Perpetuation of Health Inequities: An AI Bias Aware Framework. *Health Policy and Technology*, 12(1), p.100702. doi: https://doi.org/10.1016/j.hlpt.2022.100702.

Kent Baker, H., Pandey, N., Kumar, S. and Haldar, A. (2020). A bibliometric analysis of board diversity: Current status, development, and future research directions. *Journal of Business Research*, 108, pp.232–246. doi: https://doi.org/10.1016/j.jbusres.2019.11.025.

Khera, R., Haimovich, J., Hurley, N.C., McNamara, R., Spertus, J.A., Desai, N., Rumsfeld, J.S., Masoudi, F.A., Huang, C., Normand, S.-L., Mortazavi, B.J. and Krumholz, H.M. (2021). Use of Machine Learning Models to Predict Death After Acute Myocardial Infarction. *JAMA Cardiology*, [online] 6(6), p.633. doi: https://doi.org/10.1001/jamacardio.2021.0122.

Pesapane, F., Codari, M. and Sardanelli, F. (2018). Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. *European Radiology Experimental*, [online] 2(1). doi: https://doi.org/10.1186/s41747-018-0061-6.

Tang, X., Wu, Z., Liu, W., Tian, J. and Liu, L. (2023). Exploring effective ways to increase reliable positive samples for machine learning-based urban waterlogging susceptibility assessments. *Journal of environmental management*, 344, pp.118682–118682. doi: https://doi.org/10.1016/j.jenvman.2023.118682.

Topol, E.J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, [online] 25(1), pp.44–56. Available at: https://www.nature.com/articles/s41591-018-0300-7.