Teesside University

School of Computing, Engineering & Digital Technologies

AI ETHICS IN-COURSE ASSESSMENT

INVESTIGATING FOR GENDER-BASED BIAS IN MACHINE LEARNING MODEL (KNN ALGORITHM) DEVELOPED FOR STROKE PREDICTION

NAME: CHUKWURAH PAUL

STUDENT ID – B1267718

SUBMISSION DATE: 03/05/23

ABSTRACT

Machine learning algorithms are increasingly used in the healthcare industry to help healthcare experts make more informed decisions and improve patient care. However, this study specifically investigated a widely used algorithm (K-nearest neighbour) that could impacts millions of patients. The investigation focused on a stroke dataset to uncover gender bias, and three experiments were conducted, including gender awareness, fairness through unawareness, and splitting true and predicted values into two protected groups. The analysis evaluated the overall classification model performance and bias criteria for the two groups using three fairness criteria: Equal accuracy, Equal opportunity, and Demographic parity. This study found that the models for both groups did not satisfy any of the three fairness criteria, indicating a bias. Addressing this disparity could lead to a reduction in gender discrimination, a decrease in misdiagnosis, and an increase in trust towards AI (Artificial Intelligence) models.

Keywords: Stroke prediction, machine learning, bias, knn.

TABLE OF CONTENT

1.0 INTRODUCTION

Stroke is a serious condition that often develops when the human brain's supply of blood and oxygen is interrupted, leading to the rapid death of most brain cells (*Lo et al., 2003*). According to Johnson et al. (2016), it is the world's second leading cause of death and disability-adjusted life years. It is also responsible for an estimated 11% of all mortality. The complex nature of stroke may be well suited to the application of ML (Machine Learning) techniques, which may combine a wide range of factors and observations into a single prediction framework without the requirement for pre-programmed rules.

Every area of our life has been influenced by machine learning algorithms. Algorithms propose movies, items to purchase, potential partners, in high-stakes situations like loans (*Amitabha et al., 2002*) and employment selections (*Miranda et al., 2018, Lee et al., 2019*). To enhance diagnosis, treatment selection, and health system effectiveness, machine learning is increasingly applied in clinical care. Algorithmic decision-making has definite advantages; unlike individuals, machines do not get weary or bored (*Shai et al., 2011, Anne et al., 2010*).

Algorithms, like people, are susceptible to biases, which make their judgements "unfair" (*Julia et al., 2016, Cathy et al., 2016*). When making decisions, being fair means not showing bias or favouritism towards a person or group based on their inherent or learned characteristics. Thus, a decision-making algorithm that favours a specific group of people is unjust. Because machine-learning models get their knowledge from previously gathered data, communities that have historically been subjected to human and structural biases, often known as protected groups, are more susceptible to harm from inaccurate prediction or resource withholding.

1.1 Objective

The objective of this study is to investigate the existence of gender bias in machine learning model developed for the prediction of stroke. A dataset containing patients' information is adopted. After pre-processing the dataset, one machine learning algorithms, K-nearest neighbour is employed in the models' training process.

2.0 LITERATURE REVIEW

There are many previous works on investigation of bias in healthcare algorithms. Straw et al. (*Straw et al., 2022*) reconstructed four AI models that had been found to be more than 70% accurate at detecting liver illness from the results of blood tests in earlier studies. When they looked at how they performed by gender, they observed that they missed 44% of the liver disease cases in women compared to 23% in men. It was observed that the two algorithms regarded to be the best at diagnosing disease in patients had the greatest gender disparity, performing worse for women than men. Obermeyer et al. (*Obermeyer et al., 2019)* showed that a widely used algorithm in the healthcare, which affects millions of patients, contains significant racial prejudice: That Black patients are noticeably sicker than White patients at the same risk score, according to indicators of uncontrolled illnesses. Daneshjou et al. *(Daneshjou et al., 2022)* showed how modern dermatology AI models perform worse on Diverse Dermatology Images (DDI), with receiver operator curve area under the curve (ROC-AUC) falling by 27-36% in comparison to the models' initial test results.

2.1 Types of bias in healthcare

Implicit bias

It is an unintentional association that lead to a negative evaluation of a person based on insignificant characteristics like gender or race. (*FitzGerald et al., 2017*).

Structural Bias

Systemic elements known as structural prejudice make it difficult for some people to get healthcare. These may consist of things like a deficient healthcare system, discrimination, and poverty. Uneven health outcomes for various populations can be attributed to structural bias. (*Williams et al., 2013*)

Confirmation Bias:

It refers to the tendency of medical experts to seek and interpret facts in a way that confirms their preconceived thoughts or preconceptions about a patient's condition. This may lead to incorrect diagnoses or delayed medical care. (*Croskerry et al., 2003*)

Selection Bias:

When particular people or groups are purposefully left out of a study or analysis, selection bias arises. This may lead to generalisations and conclusions that are biased. (*Lash et al., 2009*)

3.0 METHODOLOGY

3.1 Data collection

The dataset includes 32561 observations with 15 variables representing clinical features. This data was obtained from Kaggle: stroke dataset and a summary of the dataset is shown in table 1:

NO	NAME	DESCRIPTION
1	sex	gender
2	age	age of patients
3	hypertension	hypertension status
4	heart_disease	heart disease or not
5	ever_married	marital status
6	work_type	type of patient's work
7	Residence_type	patient's residence
8	avg_glucose_level	average glucose level
9	bmi	body mass index
9	smoking_status	smoking status
10	stroke	if stroke or not

Table 1. Description of the dataset

3.2 Data Pre-processing

This stage is critical in the data analysis process for increasing the quality of data. The first step in this phase was to perform some exploratory data analysis, looking for unique values, null values, describing the dataset, and correlation matrices.

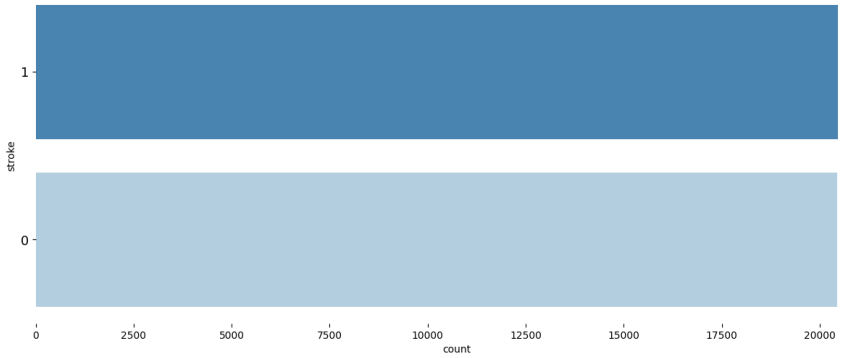

Fig 1. Stroke value count

Using correlations (relationships) between variables to view the clean data. The relationship between variables can be used to choose which variables to include in a model. A diagram illustrating the correlation between the variables is shown below:

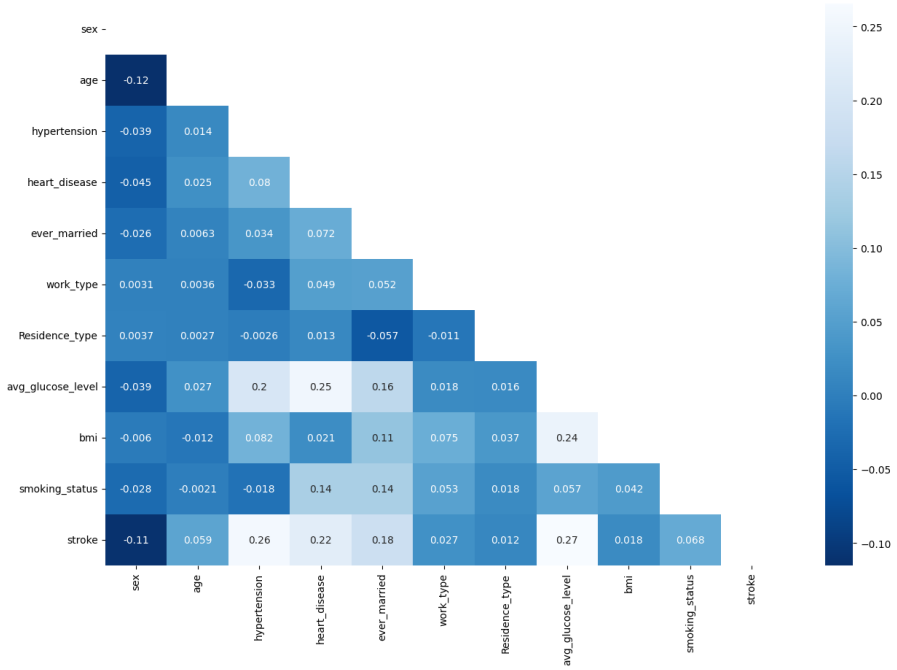

Fig 2. Correlation matrix

The data was then divided into features (X) and labels (Y). Features are the model's input and labels are the output because it is supervised learning.

3.3 Model Development

For this purpose, K-Nearest Neighbour Algorithm was used to predict stroke since it is a binary-classification problem. This model was implemented across the three experiments, in which we evaluate the performance for the overall classifier and the bias criteria for the two groups.

3.3.1 KNN

The k-nearest neighbour algorithm determines the closest neighbour of a new data item. For instance, if k=3 (where k represents the number of nearest neighbour), then the three closest neighbour are checked, and the most commonly occurring data item class is assigned to the new data item. To measure the distance between k and the new data point, we use the Euclidean distance formula. Alternatively, we can calculate the distance using the Manhattan distance formula for the KNN algorithm. (*Siddharth, 2020*)

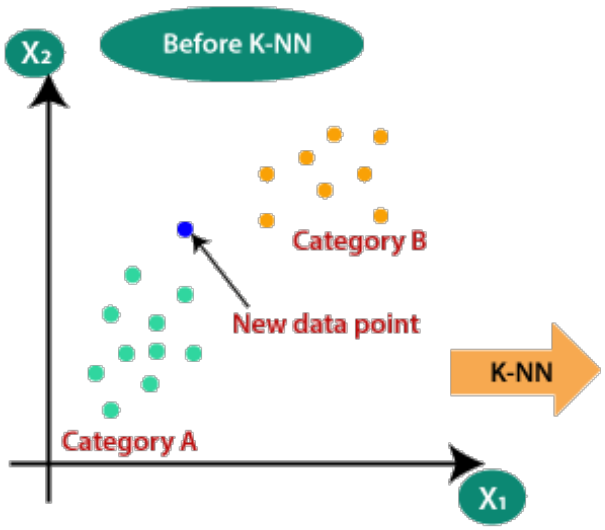

Fig 3. Before KNN (Javatpoint, 2022)

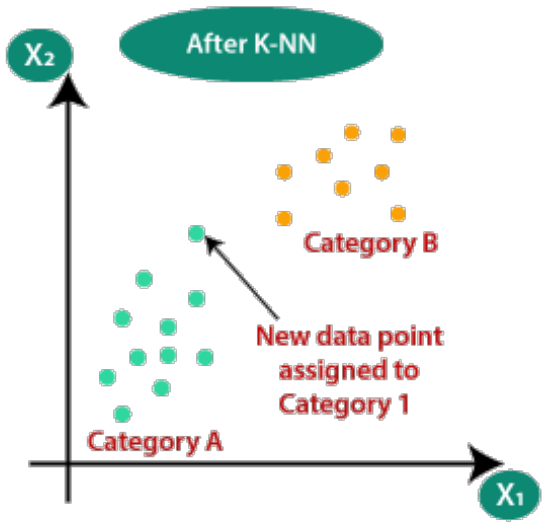

Fig 4. General Architect of KNN (Javatpoint, 2022)

Experiment 1: Gender awareness

The cleaned and pre-processed dataset was split into test and training subsets (30% and 70%, respectively), and the KNN model was applied to the dataset.

Experiment 2: Fairness through unawareness (Group unawareness)

The method used for Experiment 1 was replicated, but the sex column was removed from the train and test dataset but not the entire dataset and the model is retrained on the new features.

Experiment 3: Splitting the True and Predicted Values into two groups.

The results for experiment 1 were divided into two groups based on the actual and predicted values. In this experiment, gender served as the protected attribute; as a result, Men and Women were the two groups created. The model was retrained using the newly obtained features and labels, and the fairness criteria listed below were applied to investigate for gender-based bias.

3.4 Performance metrics for the overall classification model and the bias criteria for the two groups

Equations 1, 2, 3 present evaluation metrics for all patients while equations 1, 2, 4 present evaluation metrics for the two groups (men and women). For each evaluation metric, we look at the difference between men and women to see if there are any differences (equation 5).

Equation 1: Equal accuracy metric

The same accuracy across groups.

$$\text{Accuracy} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{True Positives (TP)} + \text{True Negatives (TN)} + \text{False Positives (FP)} + \text{False Negatives (FN)}}$$

Equation 2: Equal opportunity metric (recall)

Rate of predicted positives to actual positives.

$$\text{Equal Opportunity} = \frac{TP}{TP + FN}$$

Equation 3: Precision

The number of correct positive predictions

$$\text{Precision} = \frac{TP}{TP + FP}$$

Equation 4: Demographic parity metric

Positive outcomes should be distributed equally to each group of a protected class.

$$\text{Demographic Parity} = \frac{TP + FP}{TP + FP + TN + FN}$$

Equation 5: Sex performance disparity

Disparities in gender performance = Male group evaluation metric – Female group evaluation metric

4.0 FINDINGS AND DISCUSSION

Experiment 1: Gender Awareness

In this experiment, the model achieved an accuracy of approximately 88%, a recall rate of 97% and a precision rate of about 82% which were calculated using equations 1, 2 and 3 as outlined in section 3.4 of this report. Notably, the model recorded higher false positive rates (1336) compared to the false negative rates (194) which means the model is incorrectly predicting that more patients that do not have stroke have stroke this can lead to incorrect decisions or actions based on the model's predictions, which can be problematic in various applications. The performance metrics exhibited considerable variation when the sex column was dropped from the train and test datasets. Overall, the model demonstrated high accuracy, recall, and precision, indicating that it is capable of accurately predicting a specific category and detecting it with precision.

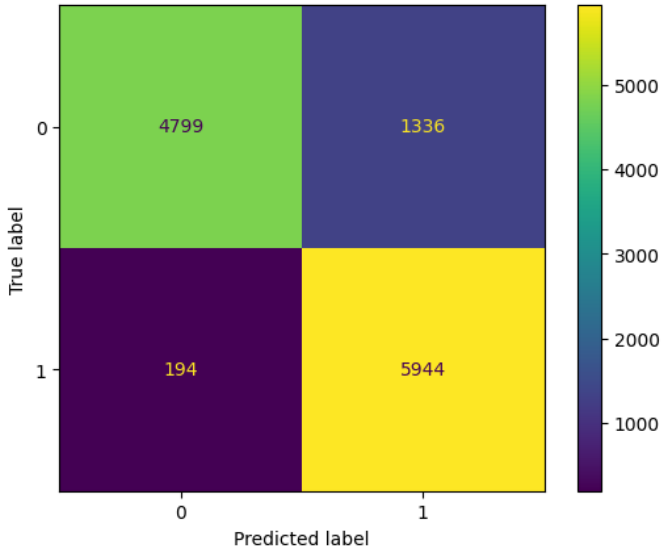

Fig 5: Confusion matrix for gender awareness

4.2 Fairness through Unawareness (Group unawareness)

The sex column was dropped from both the train and test datasets and the performance metrics were calculated. The training model yielded an accuracy score of 87%, a recall rate of 98%, and a precision rate of 81%, whereas the test model recorded an accuracy rate of 80%, a recall rate of 93%, and a precision rate of 74%. Notably, these results indicate a considerable variation from the results in experiment 1.

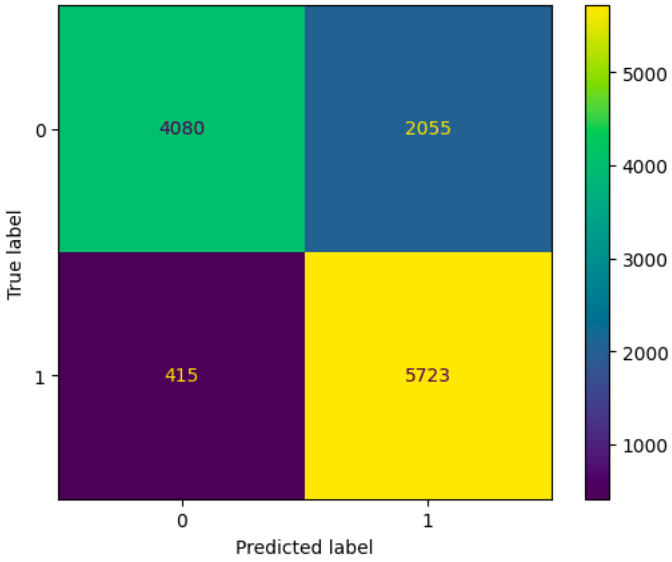

Fig 6: Confusion matrix for gender unawareness

4.3 Experiment 3 - Splitting Gender into Men and Women Group

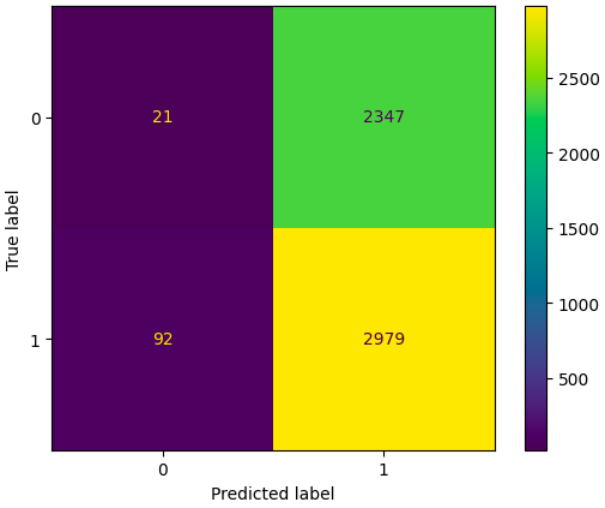

Fig 7: Confusion matrix for women's group

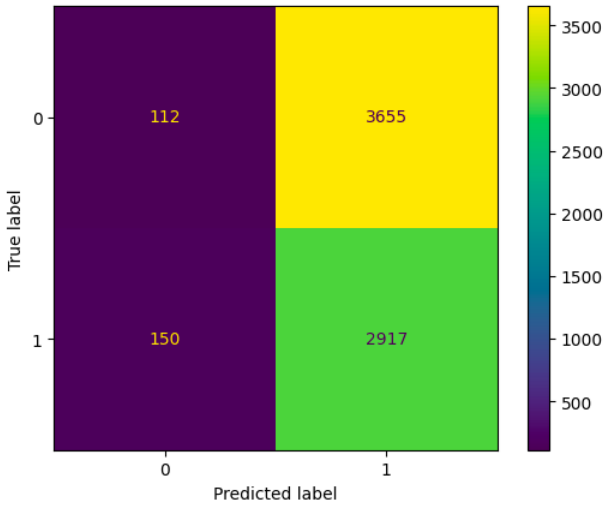

Fig 8: Confusion matrix for men's group

From the figure 6 above the True negative rate (TNR) for the women's group is 21, the False positive rate (FPR) is 2347, the False negative rate (FNR) is 92, and the True Positive rate (TPR) is 2979. While for that of the men's group the TNR is 112, the FPR is 3655, the FNR is 150 and the TPR is 2917 as seen in figure 7.

Equal Accuracy

The accuracy for both groups is calculated using equation 1 as stated in section 3.4 of this report.

$$\text{Equal accuracy for the women's group} = \frac{2979 + 21}{2979 + 21 + 2347 + 92} = 55\%$$

$$\text{Equal accuracy for the men's group} = \frac{2917 + 112}{2917 + 112 + 3655 + 150} = 44\%$$

The women's group had a mean score of 55% while the men's group had 44%. This does not satisfy the condition of equal accuracy as the difference between the two values does not equal zero which means a bias exist in the model when looked at by gender.

Equal Opportunity

The equal opportunity is calculated using equation 2 as outlined in section 3.4 of this report,

$$\text{Equal Opportunity for the women's group} = \frac{2979}{2979 + 92} = 97\%$$

$$\text{Equal Opportunity for the men's group} = \frac{2917}{2979 + 150} = 95\%$$

The women's group had a recall score of 97% and the men's group had a recall score of 95% which means each population is not receiving the same proportion of positive outcome (TRP) as required by Equal Opportunity therefore, there is a bias in the model.

Demographic Parity

The demographic parity was calculated using equation 3 in section 3.4 of this report.

$$\text{Demographic parity for the women's group} = \frac{2979 + 2347}{2979 + 21 + 2347 + 92} = 97.9\%$$

$$\text{Demographic parity for the men's group} = \frac{2979 + 3655}{2979 + 21 + 2347 + 92} = 96.2\%$$

The men's group had a mean score of 97.9% while the women's group had 96.2% which means that each segment of the protected class is not receiving the positive outcome at equal rates, therefore, does not satisfy the fairness criteria of demographic parity.

4.4 Ethical and practical implications of the results

1. Discrimination: Biased models might perpetuate gender preconceptions and lead to gender discrimination. An algorithm that forecasts higher healthcare expenditures for women, for example, may result in women being denied coverage or being charged more premiums, which is unfair and discriminatory.
2. Misdiagnosis: A biased model may not recognize certain symptoms or may overemphasize others, leading to inaccurate diagnoses.

4.5 Possible unintended consequences that could occur.

1. Reinforcing and prolonging existing inequalities: A biased algorithm has the potential to replicate and magnify existing social biases, leading to discrimination against specific groups.
2. Undermining trust in the algorithm and the organisation that employs it: When individuals learn that a model is biased, they may lose faith in the organisation that employs it. This could harm the organization's reputation and discourage customers from using its products or services.
3. Contributing to social and economic disparities: By unfairly disadvantaging certain groups, a biased machine learning model can contribute to social and economic inequities.

5.0 CONCLUSION

5.1 Knowledge Gained

My ability to consider ethical frameworks in the implementation of AI was significantly influenced by the topics presented in the Artificial Intelligence Ethics module. While exploring a variety of AI and DS (Data Science) applications, such as chatbots, medical diagnosis, fraud detection, and autonomous machines, I can confidently analyse both the risks and opportunities of using AI and DS techniques in these areas. I have also learnt that Accountability should be required of everyone, particularly of AI stakeholders, the government, businesses, and organisations.

After attempting and completing the course evaluation, I am confident that I have attained a thorough understanding of ethical concern and principles in the deployment of AI. I have developed interest in the field of AI research and ethics, and I have picked up several transferable skills that will aid in my career advancement.

5.2 Limitations and Recommendation

Limitations:

1. This study did not investigate the underlying causes of the gender bias identified in the KNN model, which may limit the potential remedies that can be used to address the bias.
2. This study only addressed three fairness criteria (Equal accuracy, Equal opportunity, and Demographic parity) and did not investigate other fairness measures or trade-offs between them.

Recommendations:

1. Investigating the underlying causes of bias in machine learning models is critical for developing effective remedies. Future studies should investigate the elements that contribute to gender bias in the KNN model and propose potential solutions to address the problem.
2. The study only focused on three fairness criteria. Future studies should investigate various fairness criteria and the trade-offs between them to provide a more elaborate evaluation of fairness in machine learning models.

REFERENCES

1. *Lo E H, Dalkara T, Moskowitz M A. Mechanisms, challenges, and opportunities in stroke. Nature reviews neuroscience, 2003, 4(5): 399-414.*
2. *Amitabha Mukerjee, Rita Biswas, Kalyanmoy Deb, and Amrit P Mathur. 2002. Multi–objective evolutionary algorithms for the risk–return trade–off in bank loan management. International Transactions in operational research 9, 5 (2002), 583–597.*
3. *Miranda Bogen and Aaron Rieke. 2018. Help wanted: an examination of hiring algorithms, equity. Technical Report. and bias. Technical report, Upturn.*
4. *Lee Cohen, Zachary C. Lipton, and Yishay Mansour. 2019. Efficient candidate screening under multiple tests and implications for fairness. arXiv:1905.11361 [cs.LG]*
5. *Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. 2011. Extraneous factors in judicial decisions. Proceedings of the National Academy of Sciences 108, 17 (2011), 6889–6892.*
6. *Anne O'Keeffe and Michael McCarthy. 2010. The Routledge handbook of corpus linguistics. Routledge.*
7. *Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: there's software used across the country to predict future criminals. And it's biased against blacks. ProPublica 2016.*
8. *Cathy O'Neil. 2016. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group, New York, NY, USA.*
9. *Straw, I., & Wu, H. (2022). Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction. BMJ Health & Care Informatics, 29(1), e100457. https://doi.org/10.1136/bmjhci-2021-100457*
10. *Obermeyer, Z., Powers, B. W., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342*
11. *Daneshjou, R., Vodrahalli, K., Novoa, R. A., Jenkins, M. M., Liang, W., Rotemberg, V., Ko, J. S., Swetter, S. M., Bailey, E. H., Gevaert, O., Mukherjee, P., Phung, M., Yekrang, K., Fong, B., Sahasrabudhe, R., Allerup, J. a. C., Okata-Karigane, U., Zou, J., & Chiou, A. S. (2022). Disparities in dermatology AI performance on a diverse, curated clinical image set. Science Advances, 8(32). https://doi.org/10.1126/sciadv.abq6147*
12. *FitzGerald, C. and Hurst, S. (2017). Implicit bias in healthcare professionals: a systematic review. BMC medical ethics, [online] 18(1). doi:https://doi.org/10.1186/s12910-017-0179-8.*
13. *Williams, D. R., & Mohammed, S. A. (2013). Racism and health I: Pathways and scientific evidence. American Behavioral Scientist, 57(8), 1152-1173.)*

14. *Croskerry, P. (2003). The importance of cognitive errors in diagnosis and strategies to minimize them. Academic Medicine, 78(8), 775-780.*
15. Siddharth Nandakumar Chikalkar. 2020. K -nearest neighbours machine learning algorithm
16. *K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint. (n.d.). www.javatpoint.com. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning*