AI Ethics and

Applications (CIS 4057N)

Investigating Bias in Lung Cancer Prediction Using Machine Learning

Word Count: 2000 without reference and table of contents

ADURA AMOO – D3622546

Table of Contents

1 INTRODUCTION

Lung cancer prediction using machine learning algorithms has shown significant promise in recent years, offering the potential to improve early detection and treatment outcomes (Altuhaifa, Win and Su, 2023). However, alongside these advancements, it's imperative to acknowledge and address the inherent biases present in such systems. In machine learning models, bias refers to systemic errors or flaws in predictions that significantly affect specific groups or individuals based on factors such as race, gender, socioeconomic status, or geographical location.

Numerous studies have demonstrated the existence of bias in healthcare algorithms, including those used for lung cancer prediction. For example, research by (Obermeyer *et al.*, 2019) revealed the existence of racial bias in a commonly utilized algorithm for allocating healthcare resources, which resulted in black patients being systematically assigned lower levels of care compared to white patients with similar health conditions. Similarly, studies by (Wiens *et al.*, 2019) and (Liu *et al.*, 2021) have identified biases in predictive models for various medical conditions, raising concerns about the fairness and equity of algorithmic decision-making in healthcare.

The implications of biased lung cancer prediction models extend beyond technical concerns to profound ethical and societal implications. Biased algorithms can exacerbate existing disparities in healthcare access and quality, perpetuating systemic inequalities and contributing to poorer health outcomes for already marginalized populations. Furthermore, the reliance on biased algorithms may erode trust in healthcare systems and undermine the credibility of medical decision-making processes.

In this report, we examine the impact of bias in lung cancer prediction using machine learning and its implications for healthcare delivery and patient outcomes. By identifying and understanding the sources and consequences of bias in predictive models, we can develop strategies to mitigate these issues and

promote more equitable and effective healthcare practices.

2 RESEACRH METHODOLOGY

There are multiple stages to this research methodology: data collection, pre-processing, visualisation, Model development, classification, Evaluation, Splitting the True and Predicted Values into Male and Female Groups, generate a confusion matrix for the two groups and apply appropriate fairness criteria.

2.1 Data Collection

Kaggle is a public repository that provides lung cancer data for this research purposes (https://www.kaggle.com/code/anaghakp/lung-cancer-prediction-logistic-regression-model/input). There are 1000 values and 26 characteristics in the dataset.

2.2 Data Pre-processing

The front end for working with the dataset was the Python environment. In order to see the dataset in a tabular format, the Data was loaded into the environment and the required libraries were loaded in pandas data frame. The describe() method is used to obtain the dataset's statistical distribution, including its mean, median, standard deviation, and so forth. The data has no null values, but the Info() function lets us see the datatype and whether the dataset has any. The seamless classification will be made possible by the conversion of the category values to numerical values. Two features index and patient id were drop because they are unique for every row and will deviate the accuracy of the model. These are shown in Figure 1-4.

	index	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	...	Fatigue	Weight Loss	Shortness of Breath	Wheezing	Swallowing Difficulty	Clubbing of Finger Nails
0	0	P1	33	1	2	4	5	4	3	2	...	3	4	2	2	3	1
1	1	P10	17	1	3	1	5	3	4	2	...	1	3	7	8	6	2
2	2	P100	35	1	4	5	6	5	5	4	...	8	7	9	2	1	4
3	3	P1000	37	1	7	7	7	7	6	7	...	4	2	3	1	4	5
4	4	P101	46	1	6	8	7	7	7	6	...	3	2	4	1	4	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	995	P995	44	1	6	7	7	7	7	6	...	5	3	2	7	8	2
996	996	P996	37	2	6	8	7	7	7	6	...	9	6	5	7	2	4
997	997	P997	25	2	4	5	6	5	5	4	...	8	7	9	2	1	4
998	998	P998	18	2	6	8	7	7	7	6	...	3	2	4	1	4	2
999	999	P999	47	1	6	5	6	5	5	4	...	8	7	9	2	1	4

1000 rows × 26 columns

Figure 1: DATASET VIEW

	index	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	...	Coughing of Blood	
count	1000.000000	1000.000000	1000.000000	1000.0000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	...	1000.000000	100
mean	499.500000	37.174000	1.402000	3.8400	4.563000	5.165000	4.840000	4.580000	4.380000	4.491000	...	4.859000	
std	288.819436	12.005493	0.490547	2.0304	2.620477	1.980833	2.107805	2.126999	1.848518	2.135528	...	2.427965	
min	0.000000	14.000000	1.000000	1.0000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...	1.000000	
25%	249.750000	27.750000	1.000000	2.0000	2.000000	4.000000	3.000000	2.000000	3.000000	2.000000	...	3.000000	
50%	499.500000	36.000000	1.000000	3.0000	5.000000	6.000000	5.000000	5.000000	4.000000	4.000000	...	4.000000	
75%	749.250000	45.000000	2.000000	6.0000	7.000000	7.000000	7.000000	7.000000	6.000000	7.000000	...	7.000000	
max	999.000000	73.000000	2.000000	8.0000	8.000000	8.000000	8.000000	7.000000	7.000000	7.000000	...	9.000000	

8 rows × 24 columns

Figure 2: SATISTICAL DISTRIBUTION OF THE DATASET

Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	Obesity	...	Fatigue	Weight Loss	Shortness of Breath	Wheezing	Swallowing Difficulty	Clubbing of Finger Nails	Frequent Cold	Dry Cough	Snoring	Level
5	4	3	2	2	4	...	3	4	2	2	3	1	2	3	4	0
5	3	4	2	2	2	...	1	3	7	8	6	2	1	7	2	1
6	5	5	4	6	7	...	8	7	9	2	1	4	6	7	2	2
7	7	6	7	7	7	...	4	2	3	1	4	5	6	7	5	2
7	7	7	6	7	7	...	3	2	4	1	4	2	4	2	3	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7	7	7	6	7	7	...	5	3	2	7	8	2	4	5	3	2
7	7	7	6	7	7	...	9	6	5	7	2	4	3	1	4	2
6	5	5	4	6	7	...	8	7	9	2	1	4	6	7	2	2
7	7	7	6	7	7	...	3	2	4	1	4	2	4	2	3	2
6	5	5	4	6	7	...	8	7	9	2	1	4	6	7	2	2

Figure 3: CONVERSION OF CATEGORICAL TO NUMERICAL VARAIBLE

	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	Obesity	...	Fatigue	Weight Loss	Shortness of Breath	Wheezing	Swallowing Difficulty
0	33	1	2	4	5	4	3	2	2	4	...	3	4	2	2	3
1	17	1	3	1	5	3	4	2	2	2	...	1	3	7	8	6
2	35	1	4	5	6	5	5	4	6	7	...	8	7	9	2	1
3	37	1	7	7	7	7	6	7	7	7	...	4	2	3	1	4
4	46	1	6	8	7	7	7	6	7	7	...	3	2	4	1	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	44	1	6	7	7	7	7	6	7	7	...	5	3	2	7	8
996	37	2	6	8	7	7	7	6	7	7	...	9	6	5	7	2
997	25	2	4	5	6	5	5	4	6	7	...	8	7	9	2	1
998	18	2	6	8	7	7	7	6	7	7		3	2	4	1	4

Figure 4: DROPPING FEATURES

2.3 VISUALIZATION

With the use of this data analysis technique, you may comprehend the underlying structure, patterns, relationships, and distributions of the data by investigating and summarising it. This is done in order to gather information and determine whether outliers are present. While eliminating outliers increases prediction performance and efficiency, in this study, the lone outlier is age, which may or may not have an impact on the data.

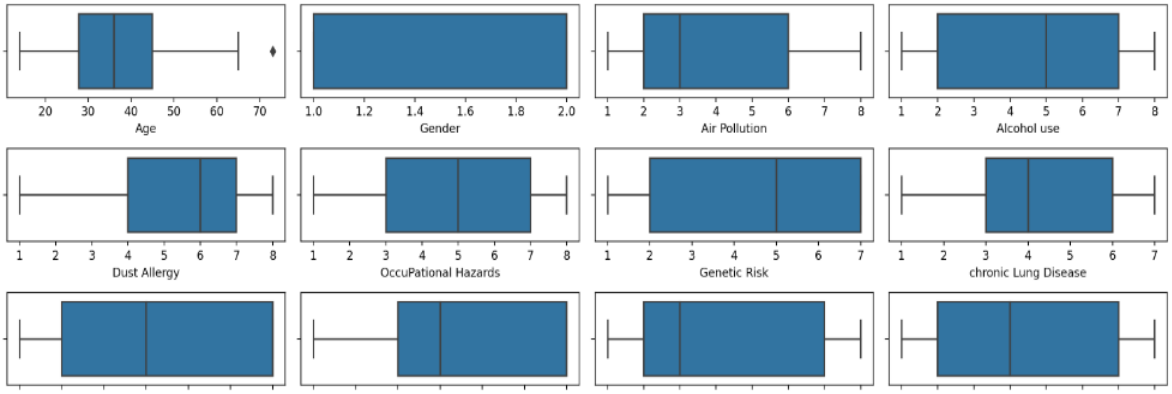


Figure 5: VIEWING OUTLIERS

The dataset was visualised using a histogram, boxplot, count plot, and correlation matrix map; some examples are shown in Figure 5-10.

2.3.1 COUNT PLOT

This seaborn visualisation tool facilitates the analysis of the relationship between the target values and the category values. The findings indicate that those that smoke are more likely to develop lung cancer than non-smokers. This distribution sample image is shown in Figure 6.

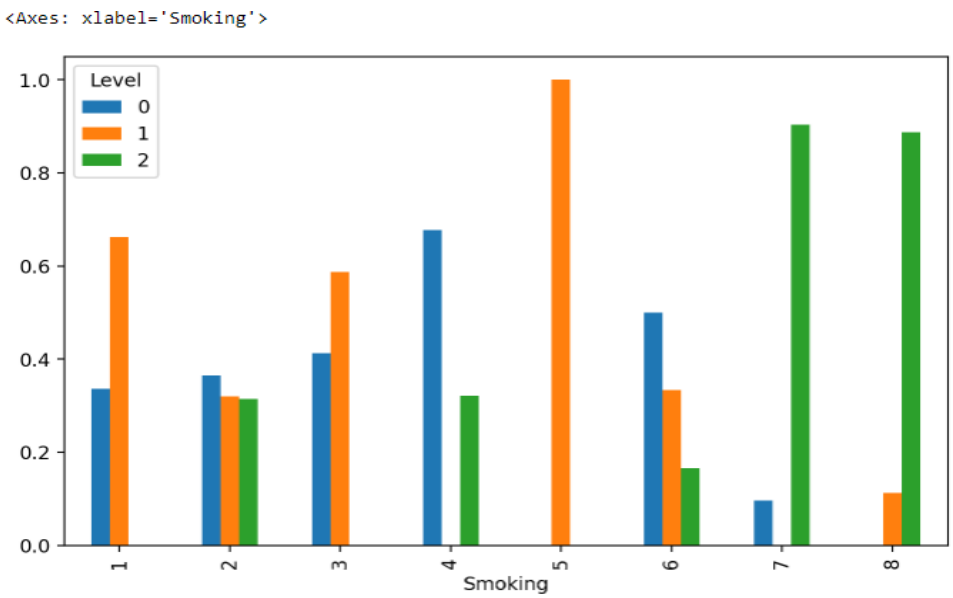


Figure 6: Using count plot to show the distribution of smokers and target variable

2.3.2 CORRELATION MATRIX

The heatmap presents correlations between variables, offering insights into their associations.

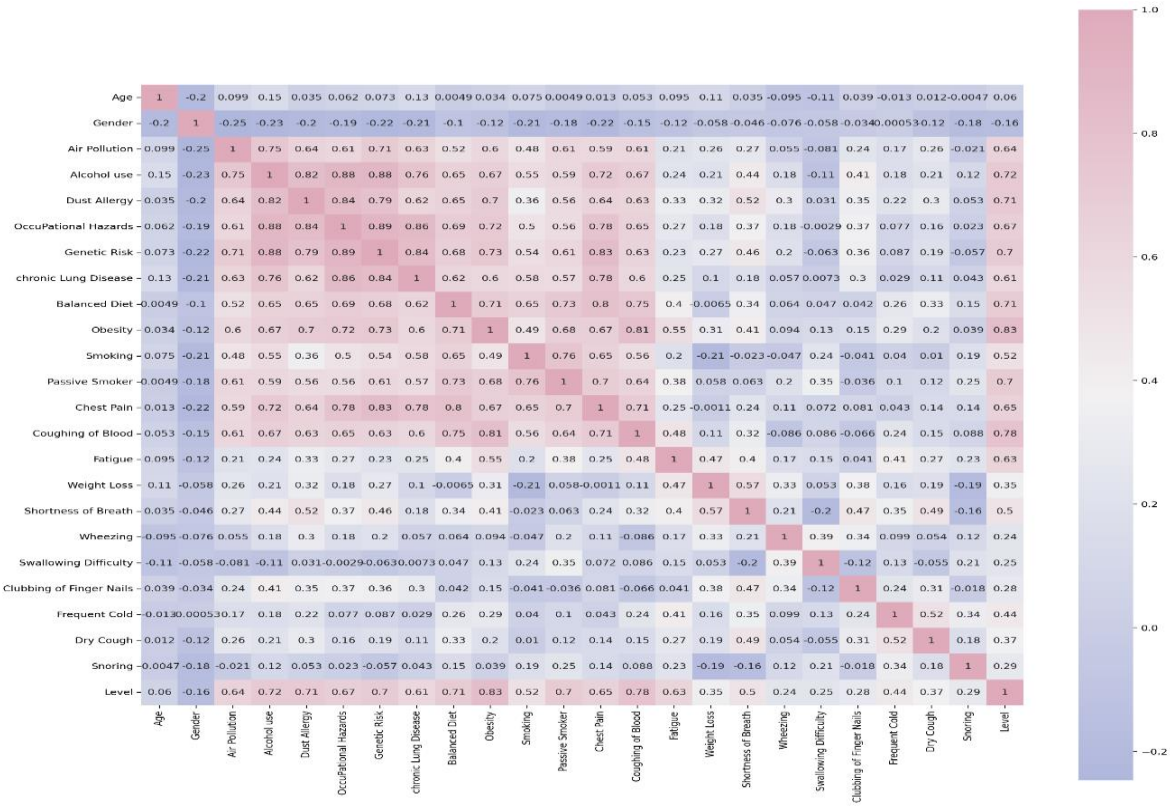


Figure 7: Using Correlation Matrix to show how Variables are Associated

2.4 Model development

Two steps are involved in building the model: data splitting and classification.

2.4.1 DATA SPLITTING

The intended variable is removed from the complete collection of data to facilitate data

separation. Subsequently, the train_test_split function is invoked to partition the data into training and testing/validation sets. Specifically, 80% of the dataset is set aside for model training, whereas the final 20% is reserved for the purpose of testing and validation.

2.4.2 CLASSIFICATION

Classification involves the process through which a model discerns patterns and relationships among input variables and output labels, enabling it to predict outcomes for new, unseen data (Jenipher and Radhika, 2020). The dataset was trained using machine learning algorithms, and predictions were generated for the test data. Gaussian Naive Bayes was use for model classification.

Gaussian Naive Bayes (GNB) is a classification algorithm rooted in probability theory. It operates under the assumption that features conform to a Gaussian (normal) distribution. Additionally, it makes the simplifying "naive" presumption that each feature is distinct , given the class label**.**(Hastie *et al.*, 2009). Gaussian Naive Bayes exhibits strong performance even with limited training data. Its model parameters, including the mean and variance for each feature and class, can be swiftly and independently estimated, facilitating rapid training, particularly advantageous for sizable datasets.

2.5 Gender-based Indexing

Through preprocessing, gender information is inferred or obtained from the dataset. This allows for the identification of male and female data points and the extraction of their corresponding indices.

2.6 Extraction of Actual Outcomes

In the methodology employed, the dataset undergoes division into training and testing sets, with the latter containing the actual outcomes for the data points. Through a preprocessing step, gender information is inferred or obtained, facilitating the identification of male and female data points and the extraction of their respective indices. Subsequently, the actual outcomes for males are extracted from the testing set using the indices of male data points, resulting in the creation of Y_test_m. Similarly, utilizing the indices of female data points, the actual outcomes for females are extracted, yielding Y_test_f. These two arrays represent the actual outcomes for males and females, respectively, enabling a gender-specific evaluation of model performance.

2.7 Extraction of Predicted Outcome

For the extraction of predicted outcomes, the methodology involves utilizing the indices of male and female data points to segregate the predicted outcomes obtained from the model (Y_predict). Specifically, for the male group, the indices corresponding to male data points are employed to subset the predicted outcomes, yielding

Y_predict_m, which embodies the predicted outcomes for males. Similarly, the indices associated with female data points are utilized to subset the predicted outcomes, resulting in Y_predict_f, representing the predicted outcomes for females. This process ensures the separation of predicted outcomes by gender, enabling a targeted evaluation of model performance within each demographic subgroup.

2.8 Conversion To Binary Class

In the process of converting the three-class problem to a binary classification task, one class is designated as representing the positive outcome, such as a high risk of having lung cancer. In this case, class 2, which signifies high risk, is mapped to the positive outcome Class1(Medium Risk). Conversely, another class is considered the negative outcome(Class 0), representing low risk, and is left unchanged. By simplifying the problem to binary classification, we focus on distinguishing between individuals at high risk and those at low risk, facilitating clearer decision-making in identifying potential cases of lung cancer. This transformation enhances interpretability and enables the application of binary classification algorithms effectively in investigating bias**.**

2.9 Apply Appropriate Fairness Criteria

In analyzing the predictive performance of a model across gender groups, it's crucial to assess and address any disparities to ensure fairness. In this scenario, the accuracy for men is notably higher than for women, indicating potential gender-based discrepancies in predictive outcomes. To promote fairness, three key fairness criteria are applied: equal accuracy, demographic parity, and equal opportunity.
Equal accuracy ensures that the model performs equally well for both genders. Adjustments are made to the model or dataset to mitigate disparities, such as retraining the model or adjusting decision thresholds to balance predictive performance between genders.

Demographic parity aims to ensure that predictions are independent of the protected attribute (gender). This criterion is achieved by adjusting decision thresholds for each gender group to ensure similar proportions of positive predictions, regardless of gender. Equal opportunity focuses on fairness in true positive rates (recall) across gender groups. Adjustments are made to the model to balance recall rates between men and women, either by modifying model parameters or applying post-processing techniques to equalize true positive rates.

3 FINDINGS & DISCUSSION

In this study, Gaussian Naive Bayes was able to correctly classify 175 cases as having low, medium or high risk of having lung cancer, 25 misclassification and type I error with 12 instances.

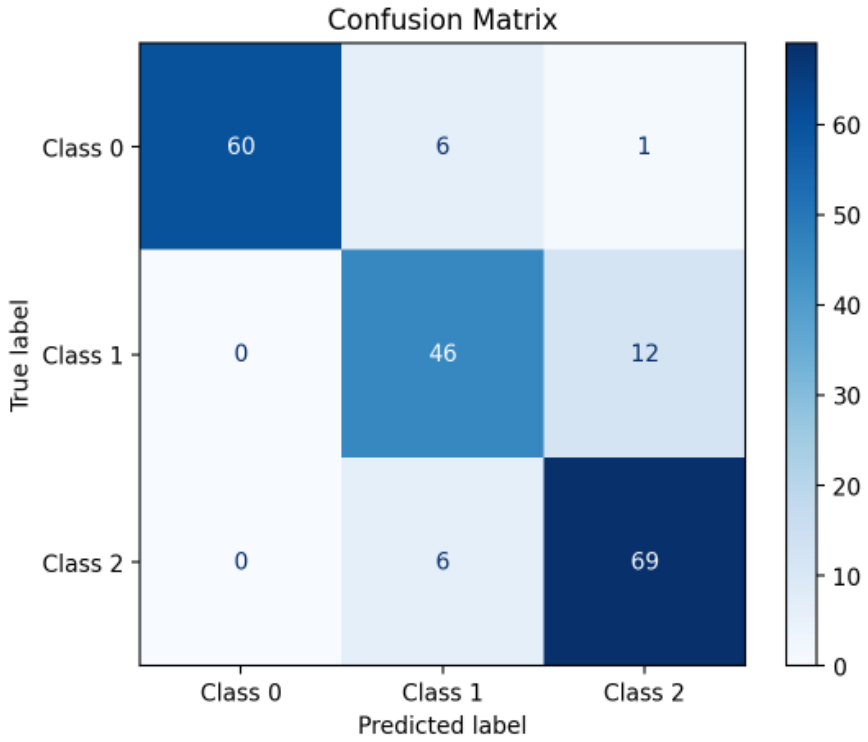
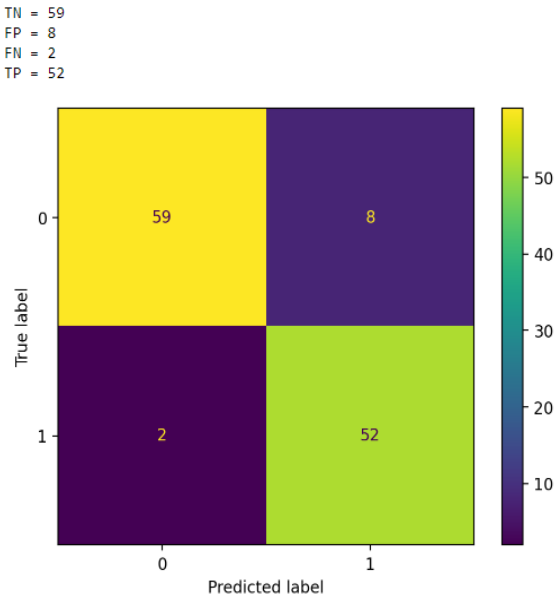


Figure 8: Gaussian Naïve Bayes Confusion Matrix

Upon analyzing the confusion matrices in the binary class for male and female classes under different fairness criteria, notable discrepancies emerge across various metrics. Under the equal accuracy criterion, which assumes that predictive accuracy should be equal across groups, similarity in accuracy is observed, where females exhibit a similar accuracy compared to males. Moving to the demographic parity criterion, which requires equal positive prediction rates regardless of the true label, the analysis reveals similarity, in both Class, indicating that demographic parity is achieved. Finally, under the equal opportunity criterion, aiming for equal true positive rates across groups for the positive class, similarity persist, notably in Class 1 and Class 2, where females exhibit same true positive rates in contrast to males. These findings underscore the significant of addressing fairness concerns in machine learning models, particularly regarding gender disparities. Further adjustments may be necessary to mitigate these discrepancies and ensure equitable outcomes across gender groups. Ongoing monitoring and refinement of the model are essential to promote fairness and build trust in the predictive capabilities of the system.

Figure 9-10 presents this.



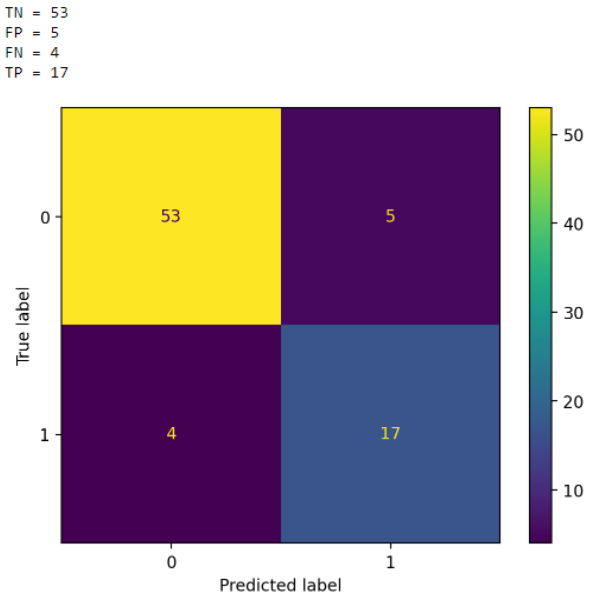


Figure 9: male confusion matrix

Figure 10: female confusion matrix

Table 1: Model Performance Comparison Across Gender Group

Performance Metric	Male	Female	Potential Bias Indication
Accuracy	88.61%	88.61%	No bias indicated
Recall	80.09%	80.09%	No bias indicated
Positive Rate	81%	81%	No bias indicated

Potential Bais indication explains where the bias leans towards based on the performance metrics provided. The positive rate being the same for both groups suggests that the model predicts positive outcomes equally for both males and females. However, the same accuracy and recall for male and female indicate that the model is more accurate and has same true positive rate for male and female, which could be considered as no bias against both genders.

The implication of a scenario when a lung cancer prediction model becomes biased towards males raises significant ethical and healthcare concerns. Such bias perpetuates

gender-based healthcare disparities, potentially leading to delayed diagnosis and unequal treatment access for females. It erodes trust in healthcare systems, undermines ethical principles of fairness and equity, and may have legal ramifications. Addressing bias requires transparent evaluation of model performance across genders, implementation of fairness-aware machine learning techniques, and collaboration among healthcare professionals, data scientists, and policymakers. Mitigating bias ensures equitable healthcare access and upholds principles of fairness and justice in healthcare delivery.

In conclusion, the performance of the model is the same between male and female groups, suggesting that the model meet the criteria for equal accuracy, demographic parity, or equal opportunity. This could help avoid fairness issues, such as bias against a particular gender.

REFERENCES

1. Altuhaifa, F.A., Win, K.T. and Su, G. (2023) 'Predicting lung cancer survival based on clinical data using machine learning: A review', *Computers in Biology and Medicine*, , pp. 107338.
2. Hastie, T. *et al.* (2009) *The elements of statistical learning: data mining, inference, and prediction.* Springer.
3. Jenipher, V.N. and Radhika, S. (2020) 'A study on early prediction of lung cancer using machine learning techniques', *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*. IEEE
4. Liu, W. *et al.* (2021) 'Demonstration and mitigation of spatial sampling bias for machine-learning predictions', *SPE Reservoir Evaluation & Engineering*, 24(01), pp. 262-274.
5. Obermeyer, Z. *et al.* (2019) 'Dissecting racial bias in an algorithm used to manage the health of populations', *Science*, 366(6464), pp. 447-453.
6. Wiens, J. *et al.* (2019) 'Do no harm: a roadmap for responsible machine learning for health care', *Nature Medicine*, 25(9), pp. 1337-1340.