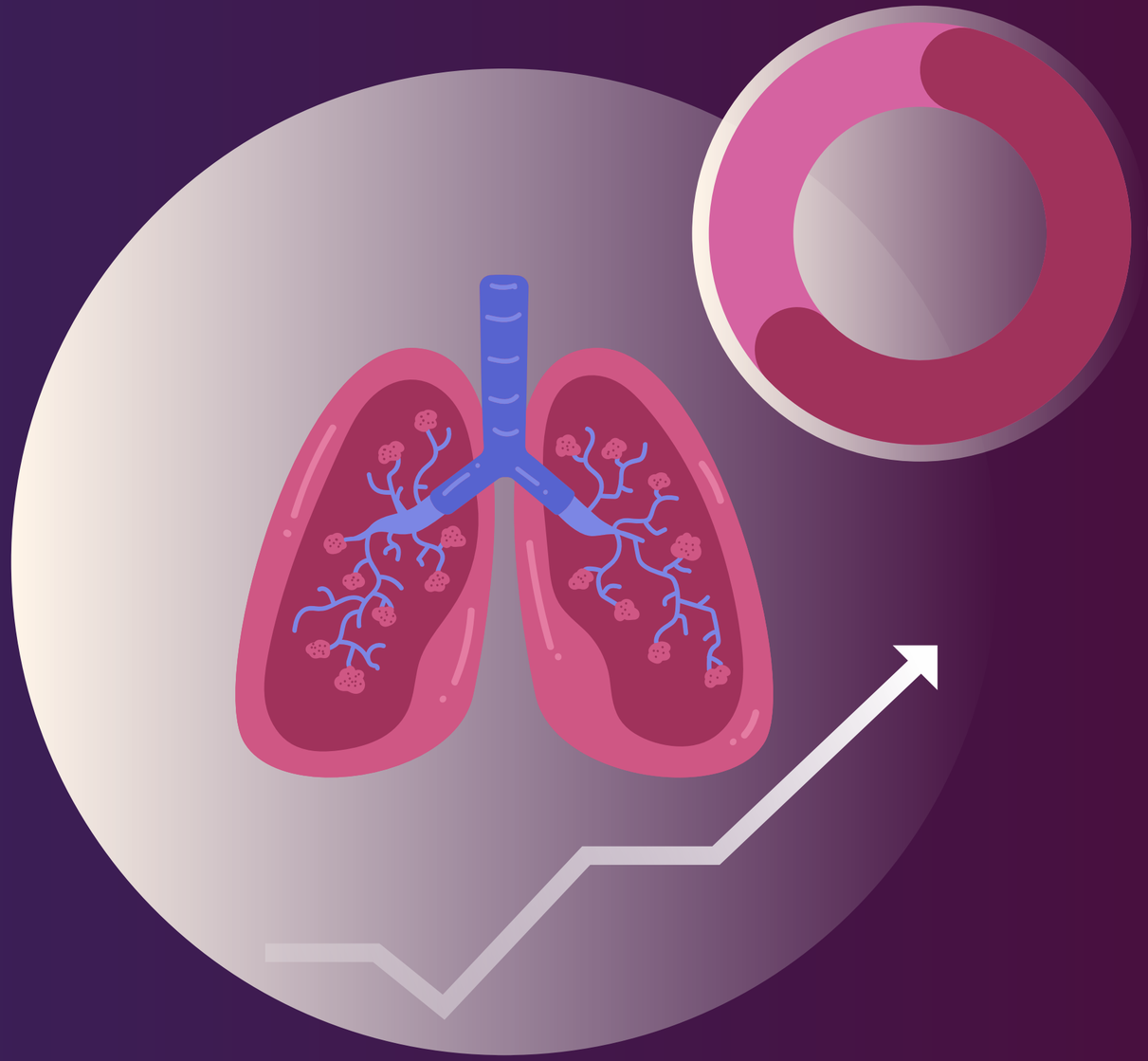
SARA SALMÓN FERNÁNDEZ

LUNG CANCER PREDICTION MODEL

IMPLEMENTATION OF A PREDICTIVE SYSTEM FOR EARLY DETECTION OF LUNG CANCER



LUNG CANCER AND ITS NUMBERS

1.8M
DEATHS

Highest number of deaths from cancer worldwide

2.5M
NEW CASES EACH YEAR

Most frequently diagnosed tumor in the world

1
CAUSE OF DEATH IN SPAIN

Main cause of death from cancer in Spain,

MORE THAN 70% OF PATIENTS ARE IN AN **ADVANCED STAGE** OF THE DISEASE AT THE TIME OF **DIAGNOSIS**

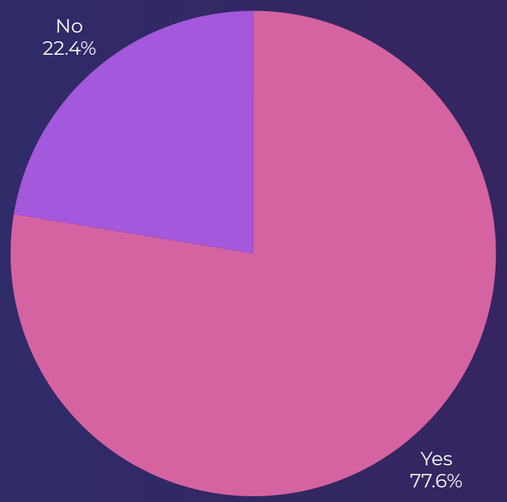
IMPROVES DIAGNOSIS SPEED, COST AND ACCESSIBILITY

LUNG AI PREDICTIVE MODEL

HELP IDENTIFY HIGH-RISK PATIENTS WITH HIGH ACCURACY

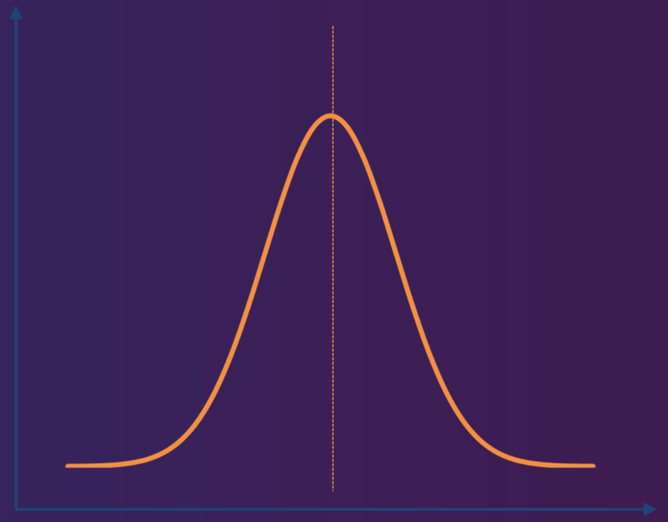
EXPLORATORY DATA ANALYSIS (EDA)

TARGET VARIABLE: CANCER



IMBALANCED DATA

INFERENTIAL STATISTICAL METHODS



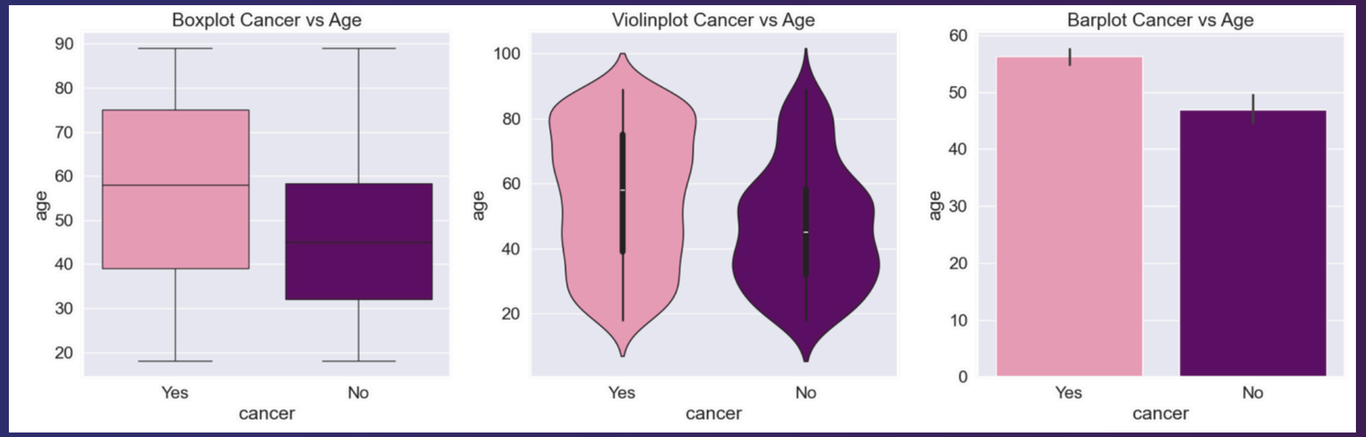
- Chi2
- T-Student
- Bar chart / Heatmaps
- Boxplots / Violin plots

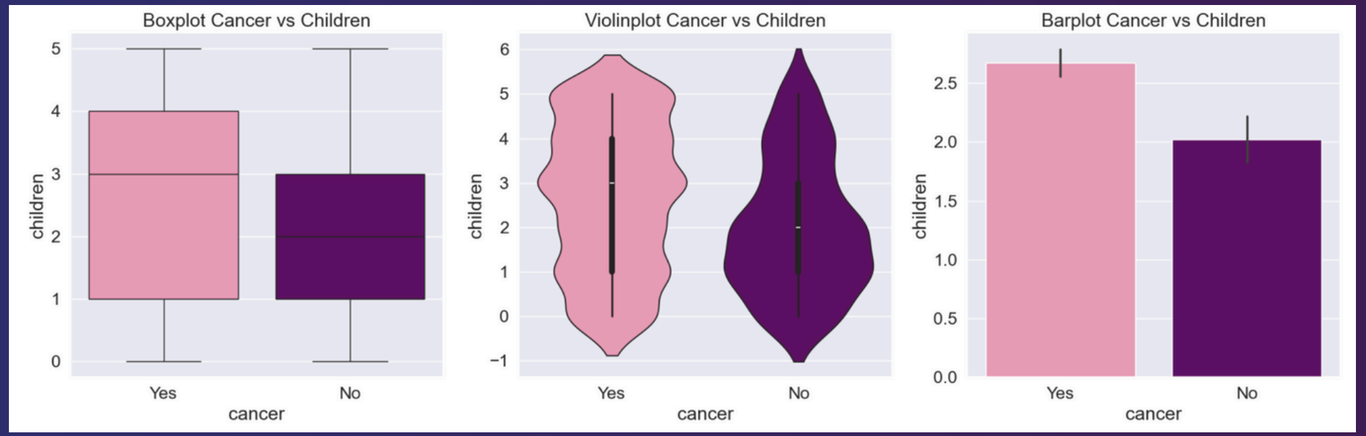
IMPORTANT VARIABLES

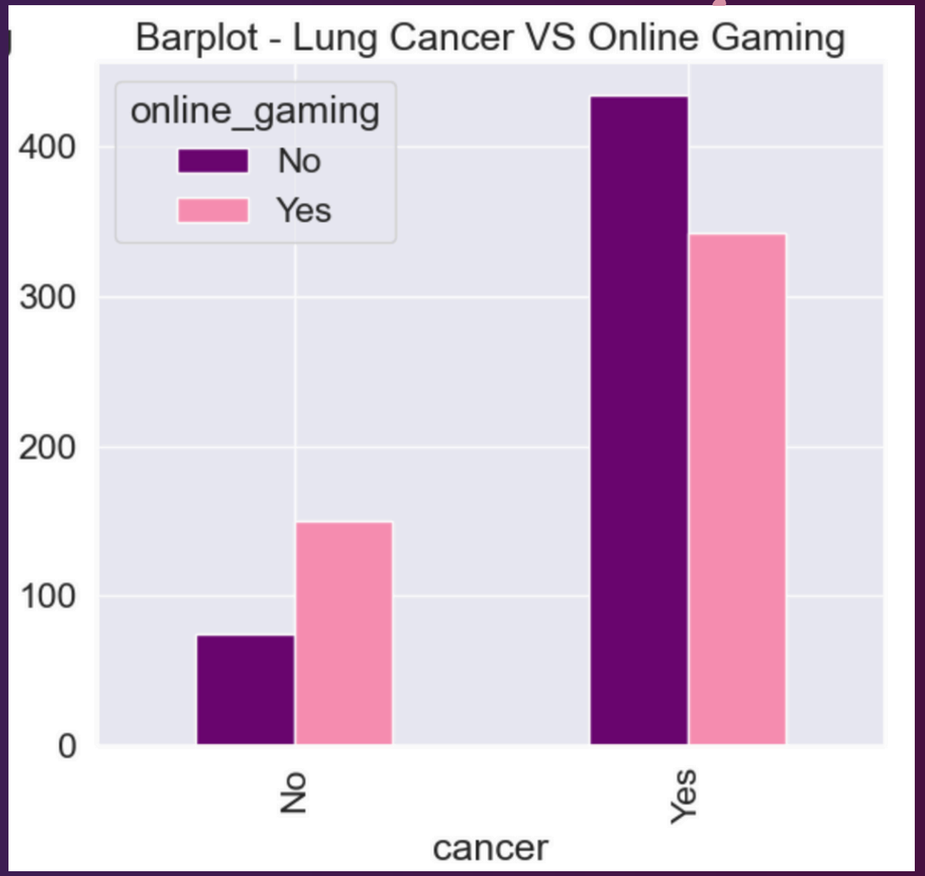
FOR LUNG CANCER PREDICITION

- AGE -RELATED
- SMOKER
- INCOME LEVEL

AGE - RELATED EDA



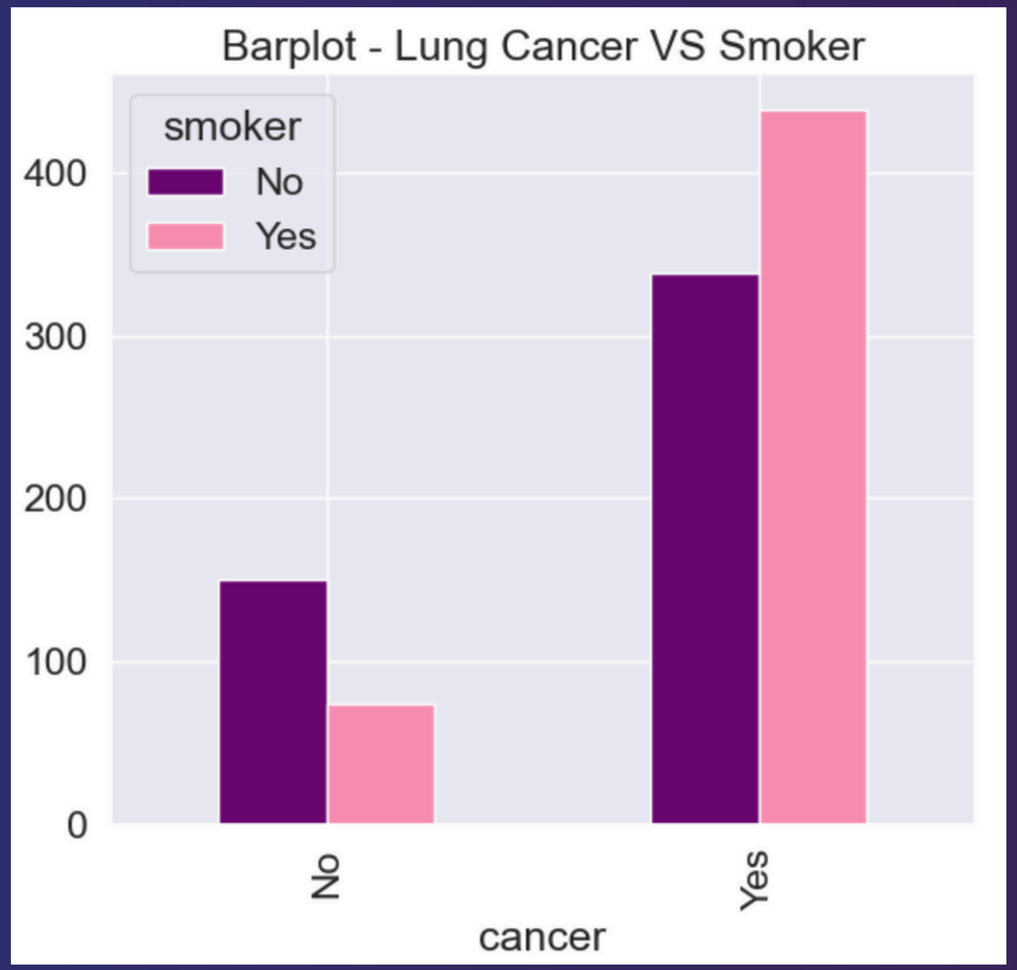


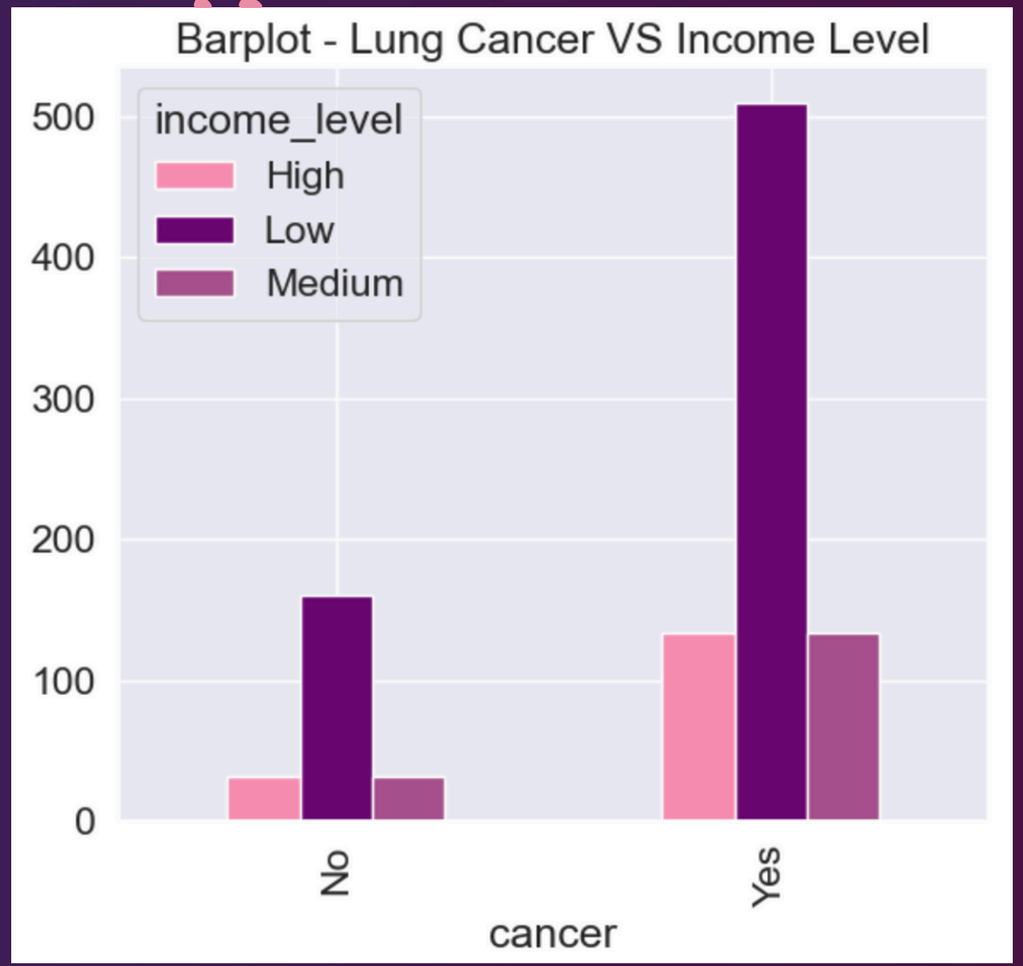


SMOKER

EDA

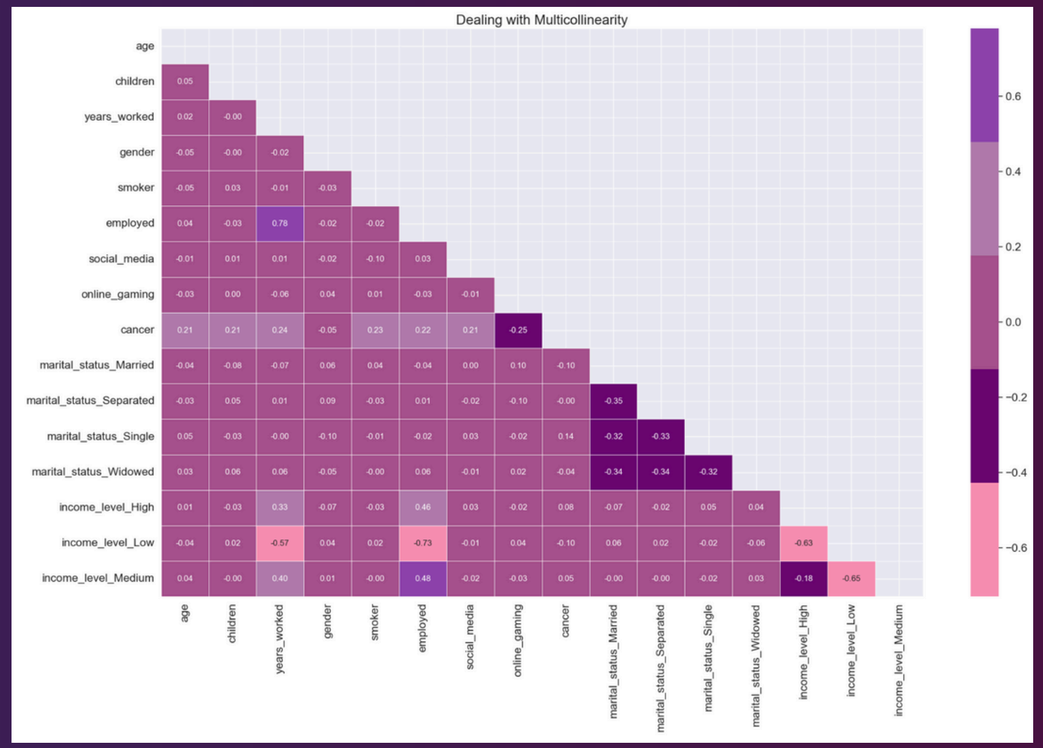
INCOME LEVEL





DATA PREPROCESSING

- DATA BALANCING
  - ✓ UPSAMPLING
    SMOTE
- CODING CATEGORICAL VARIABLES
  - ✓ OneHotEncoding
- FEATURE ENGINEERING
  - ✓ Multicollinearity Analysis and Column Elimination



FIRST APPROACH TO MODELING

Logistic Regression

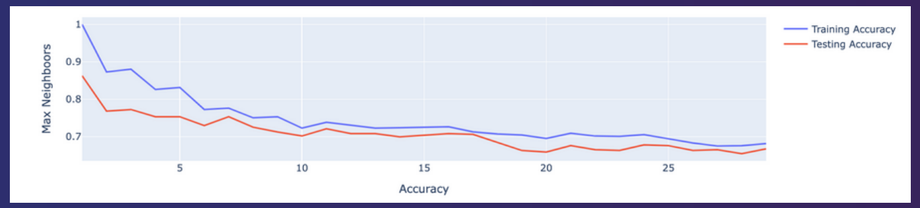
Accuracy : 0.74

K-Nearest Neighbors

Neighbors = 11 Accuracy : 0.74

Neighbors = 3 Accuracy : 0.77

Neighbors = 1 Accuracy : 0.86



Hyperparameter Tunning

Decision Trees (tree-based)

Accuracy : 0.67

LazyClassifier

- ExtraTreesClassifier
- XGBClassifier
- RandomForestClassifier

FINAL MODELING

ExtraTreesClassifier

Accuracy : 0.91

Accuracy MinMaxScaler : 0.91

Accuracy StandardScaler : 0.91

XGBClassifier

Accuracy : 0.89

RandomForestClassifier

Accuracy : 0.88

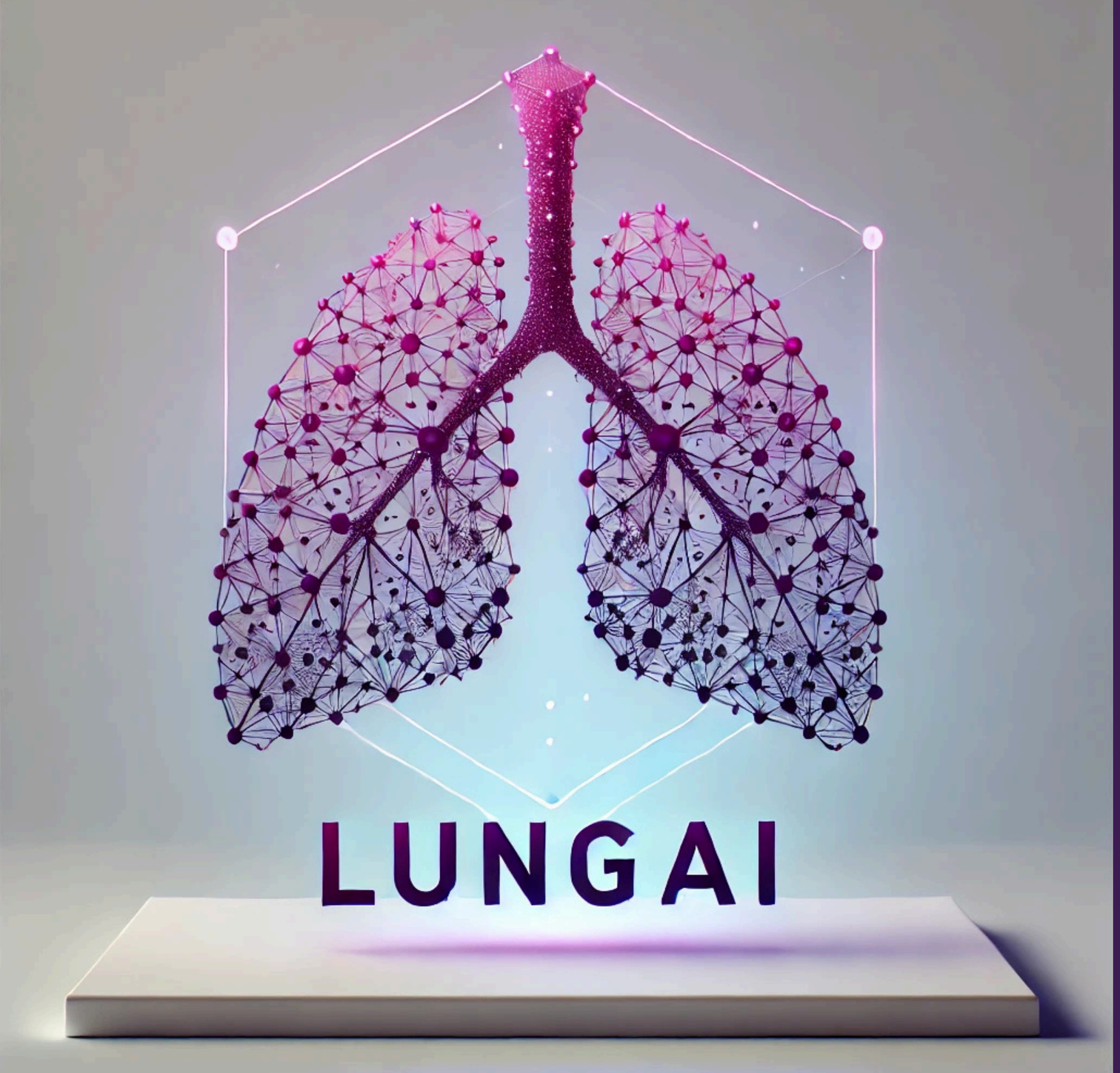
LUNG AI MODEL CONCLUSIONS

- **COMPLEMENT** TO TRADITIONAL DIAGNOSIS

- IMPROVES **SPEED AND COST** OF DIAGNOSIS

- **91% MODEL ACCURACY**

- **NEXT STEP:** USE **LUNG CT IMAGES WITH COMPUTER VISION TECHNIQUES** TO DETECT VISUAL PATTERNS OF CANCER



THANK YOU!

PATIENT DATA IS THE KEY FOR BETTER PREVENTION AND DIAGNOSIS IN HEALTHCARE!

QUESTIONS?