# Final Presentation for 3 minutes

# Slide 1

Good Morning! Everyone. Today, I will be presenting my project on 'Predicting Cancer Subtypes Using Gene Expression Data.' This represents a critical step toward precision medicine by leveraging machine learning models for robust classification."

# Slide 2

"Understanding cancer subtypes at the molecular level is pivotal for personalized treatment strategies, and this project aims to bridge the gap using advanced computational techniques.

The early detection and precise classification of cancer subtypes are essential for effective treatment. Traditional methods face challenges with the complexity of gene expression data."

"We used data from the TCGA Pan-Cancer Analysis Project, comprising 801 patient samples and over 20,000 gene expression features, covering subtypes such as BRCA, KIRC, COAD, LUAD, and PRAD.

Our objectives were to handle class imbalance, reduce dimensionality for computational efficiency, and identify significant gene features for meaningful predictions.

We ensured the dataset was clean, removed non-numeric columns, and checked for missing values."

Address class imbalance: "To mitigate bias, we applied SMOTE, generating synthetic samples for underrepresented classes like KIRC and COAD."

"This preprocessing step ensured fair representation of all subtypes, a critical foundation for building robust models.

We implemented six diverse algorithms, from Multi Logistic Regression to advanced models like XGBoost and ANN, to assess performance."

"Key metrics such as accuracy, precision, recall, F1-score, and balanced accuracy were used to evaluate model effectiveness."

"This comprehensive approach allowed us to uncover the strengths and limitations of each model."

# Slide 3

We used Stratified K-Fold Cross-Validation to ensure a robust and unbiased evaluation."

**Purpose of Cross-Validation:**  
"Cross-validation is crucial for robust model evaluation. By splitting the data into training and testing sets multiple times using Stratified K-Fold, we ensure that each subtype is proportionally and addresses the imbalance in our dataset, ensuring minority subtypes like KIRC and COAD are adequately represented. This prevents the model from being biased toward majority classes like BRCA during training."

"This method reduces overfitting by validating the model on diverse subsets, offering reliable performance estimates across all cancer subtypes."  
  
Accuracy was chosen to evaluate the final performance of the model on imbalanced data. Unlike standard accuracy, it gives equal importance to all subtypes, regardless of their representation."

"It computes the average recall across all classes, ensuring that minority subtypes like KIRC and COAD are not overshadowed by the majority class, BRCA."

"The confusion matrix is essential for evaluating model performance. It provides insights into true and false classifications, ensuring that minority subtypes are not overlooked. By highlighting misclassifications, it guides improvements in feature selection and model tuning. Additionally, it allows us to compare algorithms consistently, reinforcing the reliability of our results."

* Clinical implications:  
  "Together, these techniques ensure the model is not only accurate but also fair and clinically applicable."

**Why It Fits:**

This placement aligns the **confusion matrix** with metrics and validation techniques.

* It bridges the transition between evaluation methods (cross-validation and balanced accuracy) and their practical implementation (confusion matrix).

While Stratified K-Fold Cross-Validation ensures a balanced and fair training-testing process, balanced accuracy provides a critical measure of how well the model performs on all subtypes, especially underrepresented ones, in the final evaluation.