**Slide 1: Title Slide – Project Overview**

Welcome everyone!

Today we’re presenting our capstone project on a machine learning solution that predicts the **functional impact of p53 gene mutations**, which are strongly linked to cancer development.

This was built using real-world **bioinformatics data from UCI's p53 Mutants dataset**, containing over 31,000 samples and 5,408 binary attributes per sample.

I’m Soura, and along with my teammate Tushar, we developed an end-to-end system that includes data processing, model training, and a deployed app to make predictions in real-time.

**Slide 2: Business Problem & Motivation**

Let’s understand the motivation behind this project.

In cancer genomics, one of the biggest challenges is identifying whether a specific mutation in a gene like **p53** leads to functional loss — which is a key marker in tumor development.

Traditionally, such analysis requires **biological lab experiments** which are expensive, time-consuming, and not scalable.

Imagine being able to **instantly predict** if a mutation is harmful using just the mutation vector. This could help research labs prioritize experiments, assist biotech firms in drug development, and support hospitals working on genetic diagnostics.

Our goal was to build a system that does exactly this — a machine learning-based tool to automate **mutation impact prediction**, bringing speed and scale to the field.

**Slide 3: Data Summary**

Our dataset is the well-known *p53 Mutants* dataset, sourced from the UCI Machine Learning Repository. It includes over **31,000 samples**, each representing a unique mutation of the p53 protein — a protein that plays a critical role in cancer suppression. Each mutation is encoded into a high-dimensional **binary vector** with **5,408 features**.

The task is to predict whether the mutation inactivates the protein or not. While the original problem is binary classification, we adapted its structure to simulate multi-label scenarios by building a pipeline that handles **high-dimensional structured data** and **binary labels** — just like in real-world multi-tagging systems.

**Slide 4: Our Solution – High-Level Architecture**

With over 5,000 binary attributes per mutation, working with this dataset was like finding patterns in a massive haystack.

We started by cleaning the data and handling any formatting issues. Then we used **Principal Component Analysis (PCA)** to bring down dimensionality from 5,408 to just 150 — which preserved the information while making it easier for our model to learn.

For prediction, we used the **XGBoost algorithm**, known for its speed and performance in high-dimensional data.

The final piece was making our model **interactive and user-friendly**, so we used **Streamlit** to deploy a web-based app where users can input or upload mutation vectors and get instant predictions.

**Slide 5: Demo Preview**

Let’s take a quick look at how this works.

Here’s the **Streamlit interface** we developed. It offers two main options:

1. Load a predefined sample of mutation attributes and get a predicted label (Active or Inactive), or
2. Upload a CSV with multiple rows of binary vectors to predict them all at once.

On the backend, the data flows through our trained XGBoost model which makes the classification based on patterns it learned from thousands of previous mutations.

It’s fast, clean, and replicable. This kind of functionality can easily be extended into bioinformatics dashboards, clinical tools, or drug discovery pipelines.

**Slide 6: Business Impact & Insights**

The business and scientific potential here is significant.

Researchers can use this model to **quickly pre-screen mutations** for functional impact before committing to costly lab tests. This reduces R&D cost and accelerates timelines.

Biotech companies working on personalized medicine can integrate such tools into their data pipelines to improve **genomic screening**, drug testing, or biomarker identification.

And because the model is **generic in structure**, we can easily retrain it with BRCA, TP53, or other datasets to create customized solutions across **oncology, pharmacogenomics, and genetic counseling.**

So this is not just a student project — it’s a scalable, real-world tool.

**Slide 7: Team Collaboration**

This was a truly collaborative project, and we both took on key roles at every stage.

I focused more on **data engineering and model development**, including PCA and the training process using XGBoost. I also handled the backend integration for the Streamlit app.

Tushar focused on the **app’s user experience**, model validation, and created the clean visualizations and layouts for both the application and this presentation.

We worked iteratively on GitHub and regularly met to align progress and fix issues.

**Slide 8: Challenges & Learnings**

Working with this kind of biological data posed multiple challenges.

First, the **high dimensionality** — with 5,408 binary attributes — made it difficult to train models efficiently. We tackled this with PCA, which reduced redundancy while keeping essential signal.

Second, there was a **slight class imbalance**, which we addressed using XGBoost’s tuning options like scale\_pos\_weight and early stopping.

More importantly, we learned how to work with **real-world bioinformatics data**, build a full ML pipeline, and make it usable through a real-time web app. This project brought together theory, code, and application.

**Slide 9: Conclusion**

To summarize — we tackled a complex bioinformatics problem using a machine learning approach designed for real-world impact.

We started with a **high-dimensional dataset** of over 31,000 mutation records, each represented by **5,408 binary attributes**. The goal was to predict whether a mutation results in a **functionally active or inactive** protein — a critical indicator in cancer studies.

Our solution combined **Principal Component Analysis (PCA)** for reducing dimensionality, **XGBoost** for high-performance classification, and **Streamlit** to build an interactive interface for real-time predictions.

The result is a **scalable, fast, and user-friendly tool** that can assist in data-driven genetic research or decision-making — and it's designed to be extensible to similar high-dimensional classification problems in other domains as well.