# AI & OMICS INTERNSHIP FINAL PROJECT


Transcriptomic Profiling and Machine-Learning–Based Biomarker Identification in Pancreatic Ductal Adenocarcinoma.

## GROUP MEMBERS

MUHAMMAD ADIL KAMAL KHAN

MUHAMMAD HASSAN

KHIZRA

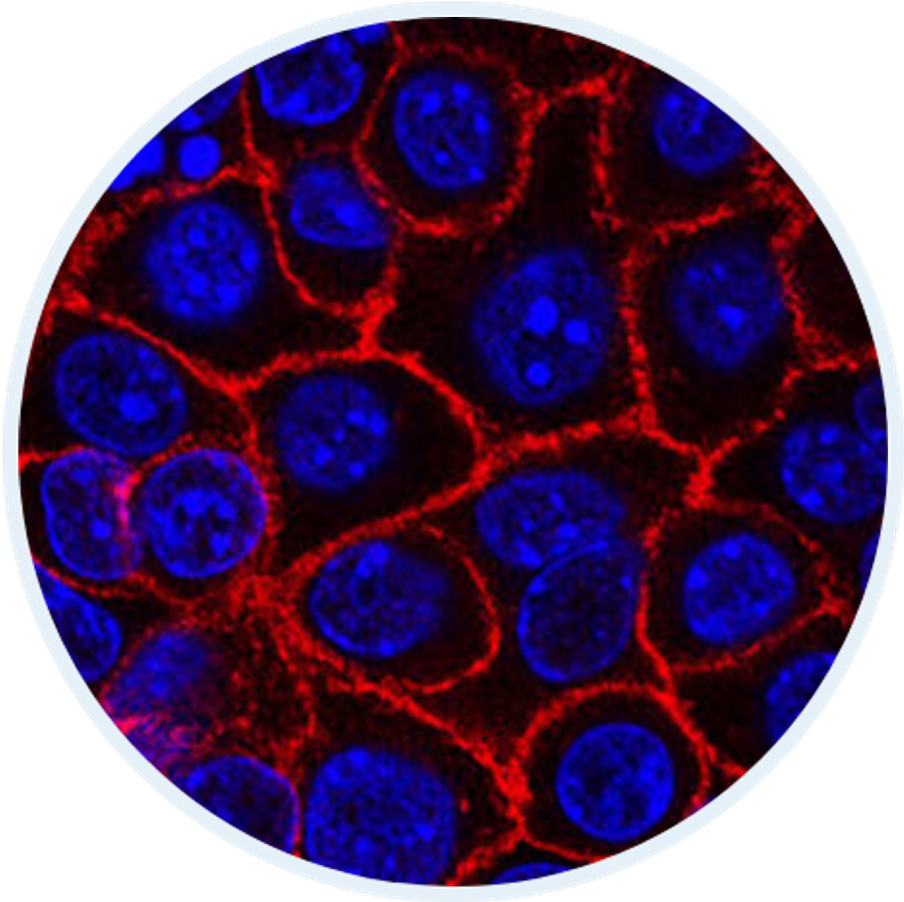
FAIZA

# THE CHALLENGE

Pancreatic Ductal Adenocarcinoma

(PDAC)

# Need for Precision Diagnostics



### Early Detection is Key

Pancreatic cancer remains one of the most lethal malignancies due to late-stage diagnosis. Identifying molecular signatures (biomarkers) from high-throughput omics data is critical for early intervention.

This project leverages the **GSE28735** dataset to build a robust AI classifier for tumor vs. healthy tissue detection.

# Study Foundation: GSE28735

## Sample Cohort

Analysis of matching pairs (45 tumor/45 normal) from 45 patients, providing a highly controlled environment for differential expression analysis.

## Technological Base

Affymetrix GeneChip Human Gene 1.0 ST Array. This platform captures the entire transcript, ensuring high resolution for biomarker detection.

# METHODOLOGY

Computational Workflow &

Processing

# The Bio-ML Pipeline

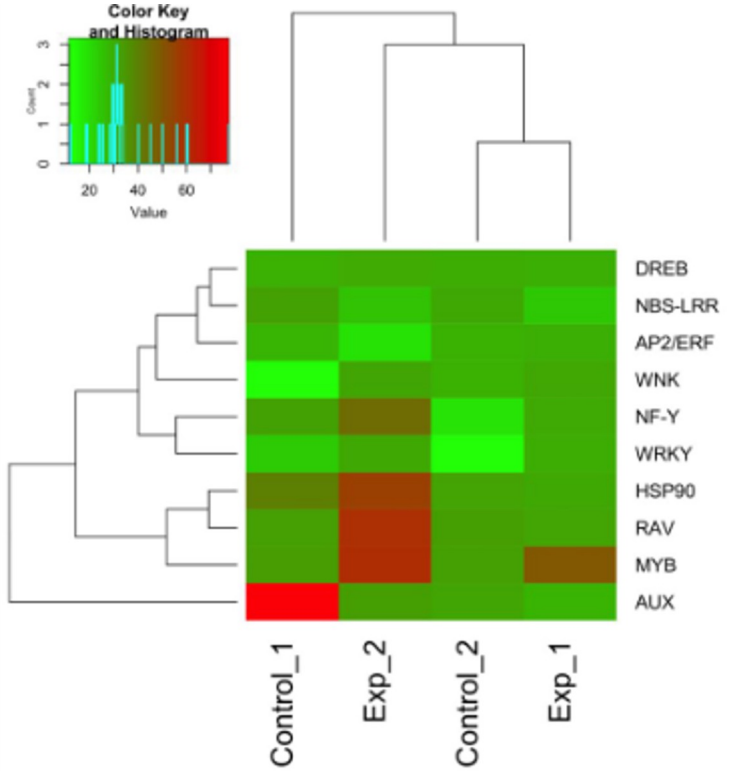
| Acquisition | Normalization | Annotation | Selection | Modeling |
| --- | --- | --- | --- | --- |
| GEOquery GSE28735 | RMA & Intensity Filtering | Probe-to-Gene Mapping | Boruta Feature Importance | Random Forest Classifier |

# Data Refinement Strategy

### Removing Noise

Raw CEL files were processed using **Robust Multi-array Average (RMA)** for normalization.

Intensity-based filtering (threshold > 3.5) ensured only highly informative transcripts were passed to the machine learning engine.



# Advanced Feature Selection

**28**

Confirmed Genes

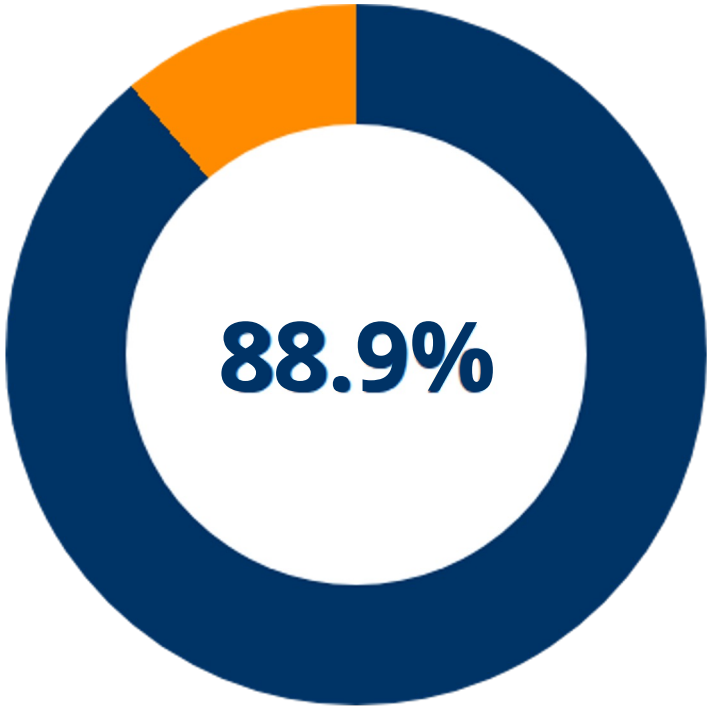
### Why Boruta Selection?

Standard selection methods often overlook relevant but non-dominant features. The **Boruta Algorithm** creates "Shadow Features" by shuffling original data to set a statistical threshold for true relevance. This identified 28 highly predictive genes for the final Random Forest model.

# Top Predictive Biomarkers


***Note:*** *Importance based on Mean Decrease Accuracy in Random Forest.*

# Model Classification Success




### Performance Metrics

The Random Forest classifier achieved a staggering **88.9% accuracy** on the independent test set.

This level of precision indicates the 28-gene signature is a robust candidate for diagnostic panel development.

# Key Project Takeaways

- **Optimized Workflow:** Integrated Bioconductor packages with advanced ML libraries (Boruta, randomForest).
- **Scientific Validity:** Identified genes like FAP and POSTN, which are known indicators of tumor stroma transformation.
- **AI Scalability:** The pipeline is adaptable to other high-throughput microarray datasets for various cancers.
- **Future Directions:** Validation via qPCR or external independent validation cohorts (e.g., TCGA).

# THANK YOU!