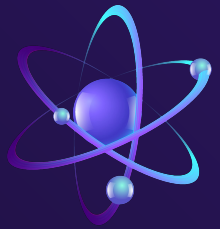
DNA Methylation Signature Analysis

Marva L.

June 13, 2022

DSI 321 Capstone

Outline



01 Introduction **The Dataset**

DNA methylation signature;
TCGA project

02 Modeling
Supervised Learning

Decision Tree &
Random Forest

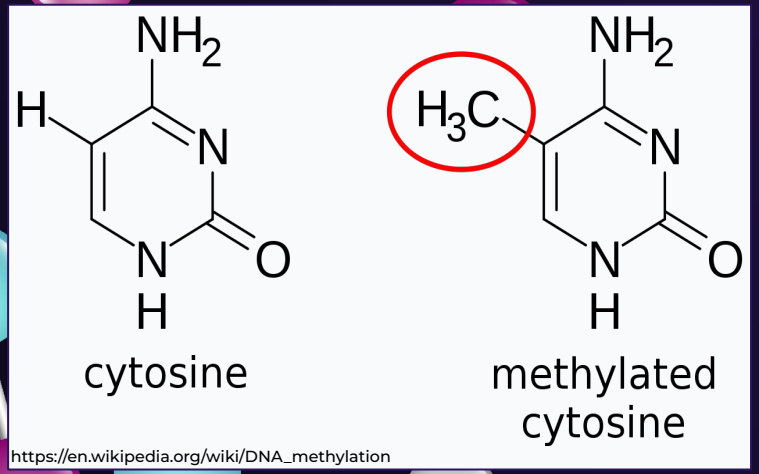
03 Sanity Check
Unsupervised Learning

tSNE

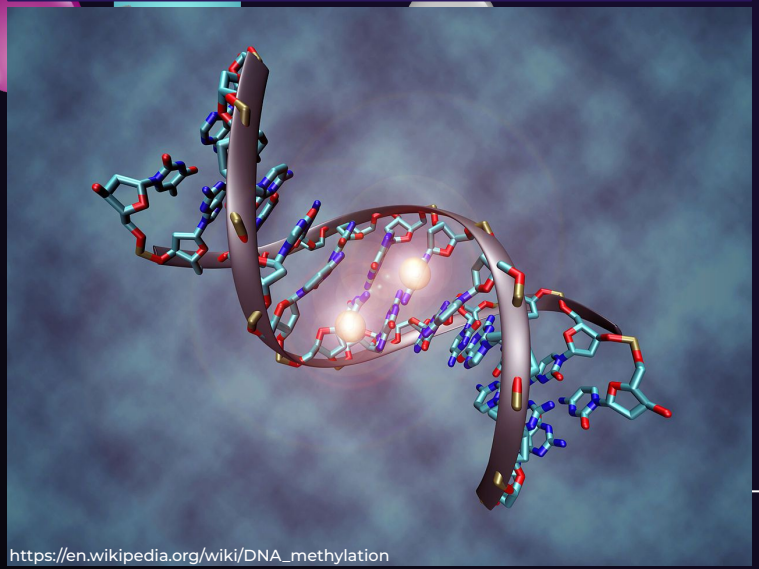
04 Conclusion &
Recommendations

01

Introduction
The Dataset



https://en.wikipedia.org/wiki/DNA_methylation

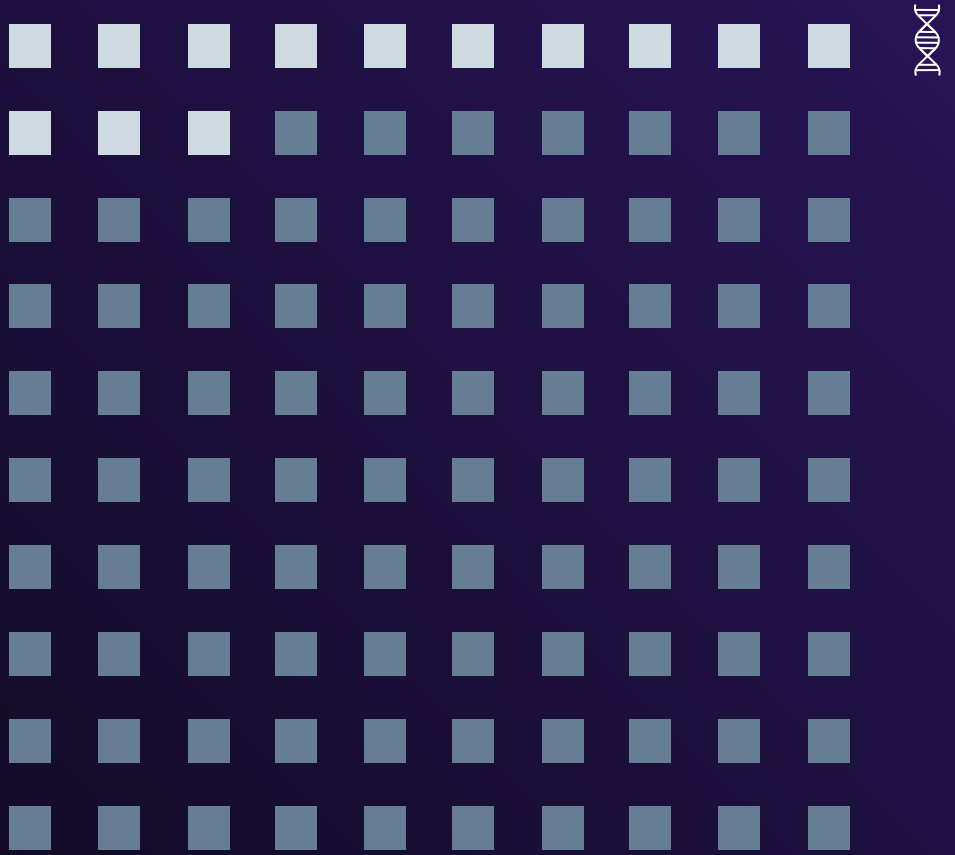

https://en.wikipedia.org/wiki/DNA_methylation

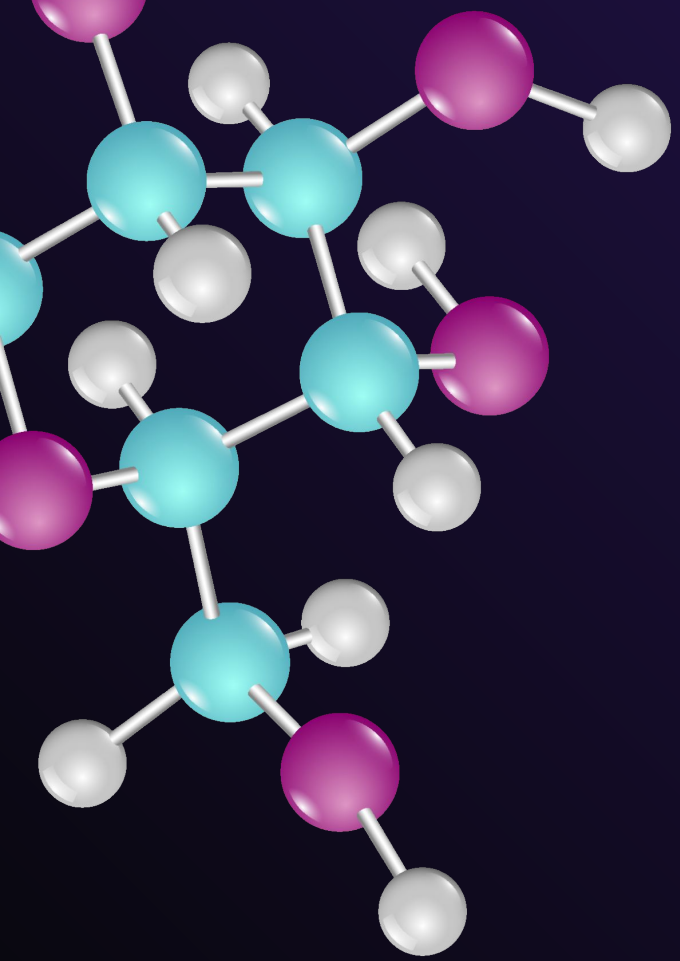
DNA Methylation Signature & Cancer

The TCGA project
https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

The Dataset
Challenges

- Imbalanced data
- Over 420k features
  - Subset of 9,000 features (2%)





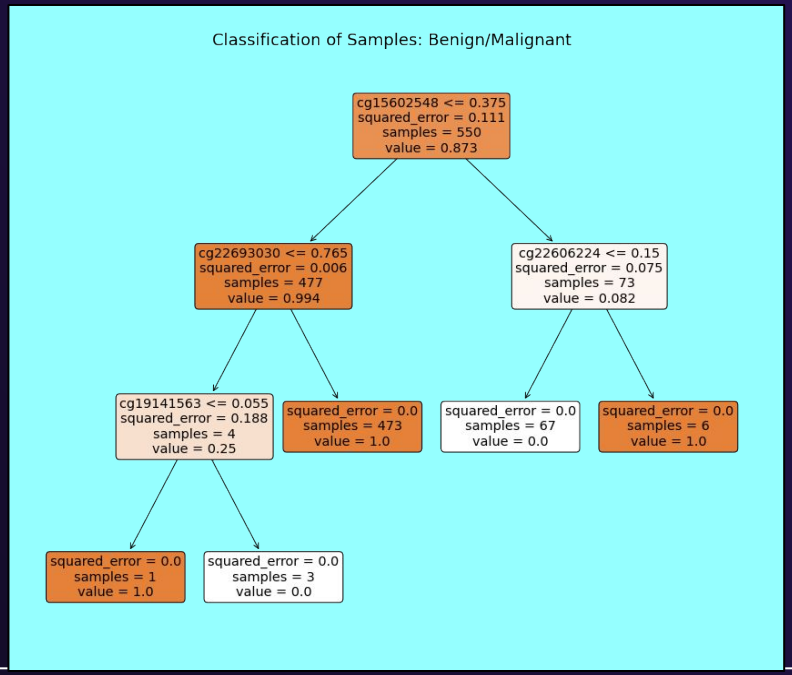
02

Supervised Learning

Decision Tree & Ensambles

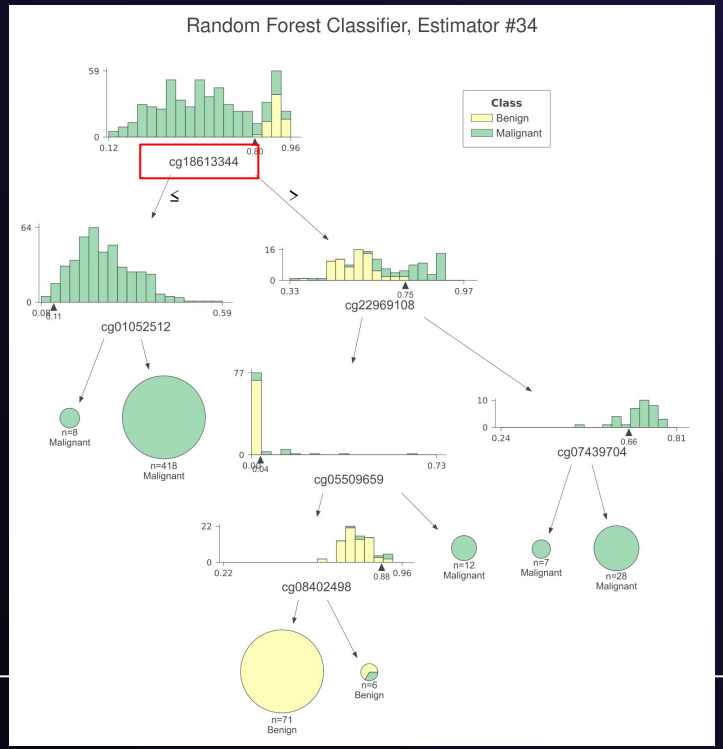
First Model
A Decision Tree

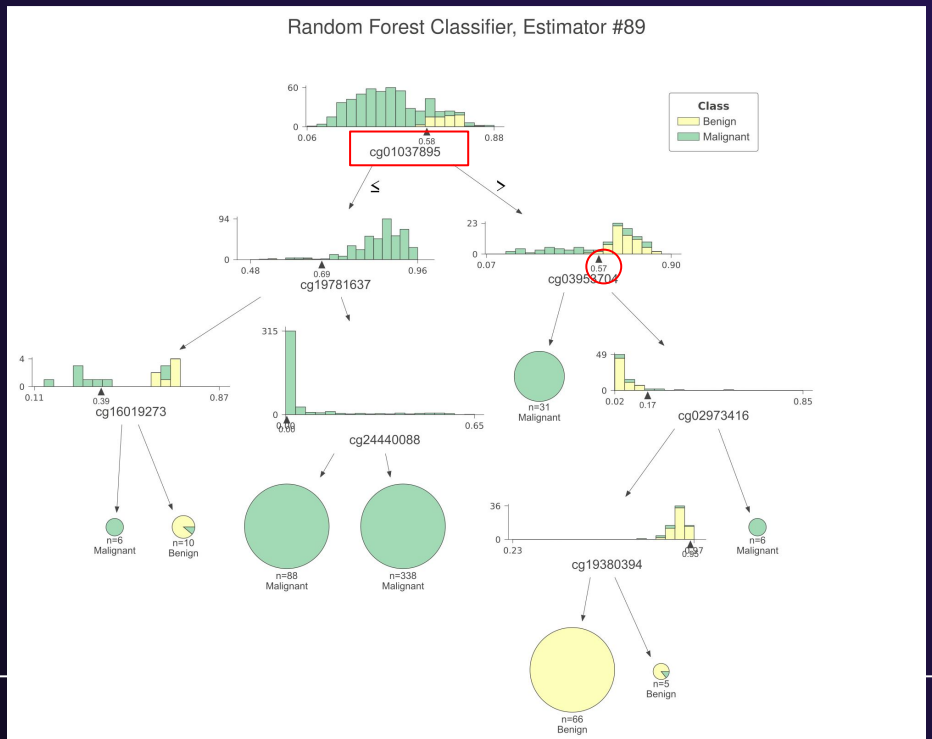
Train	1.0
Test	0.47



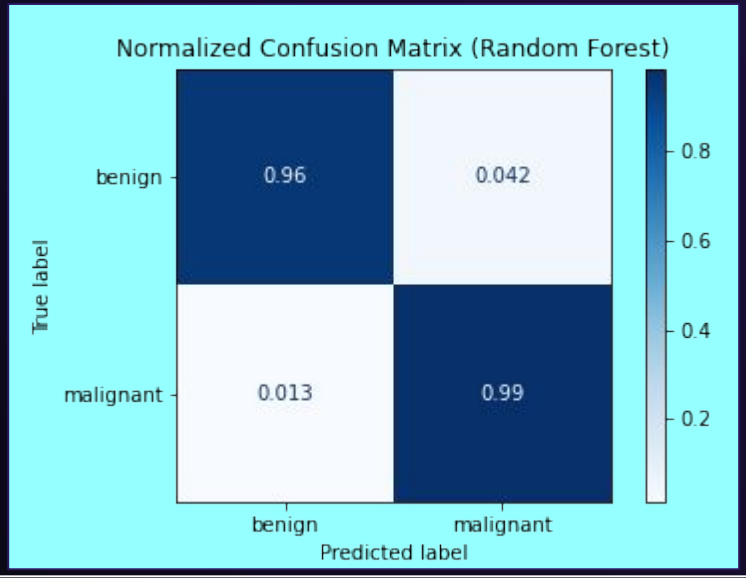
Second Model
Random Forest

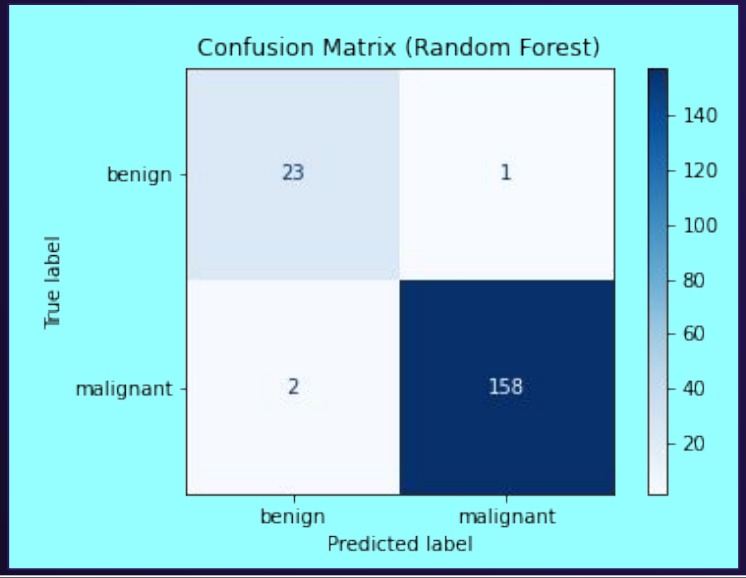
Train	0.998
Test	0.983

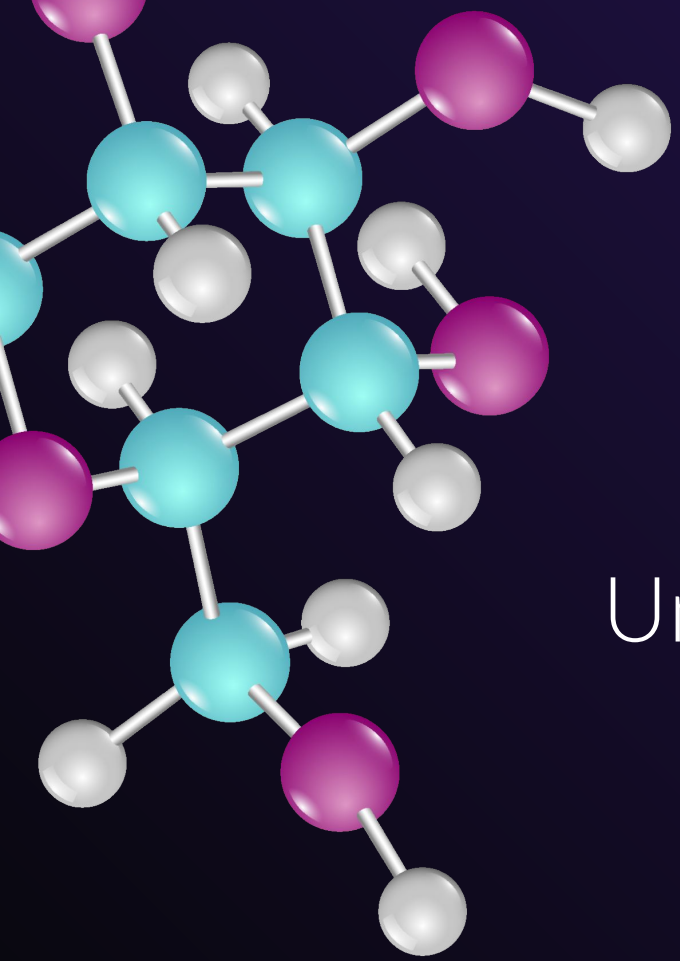




Random Forest
Confusion Matrix



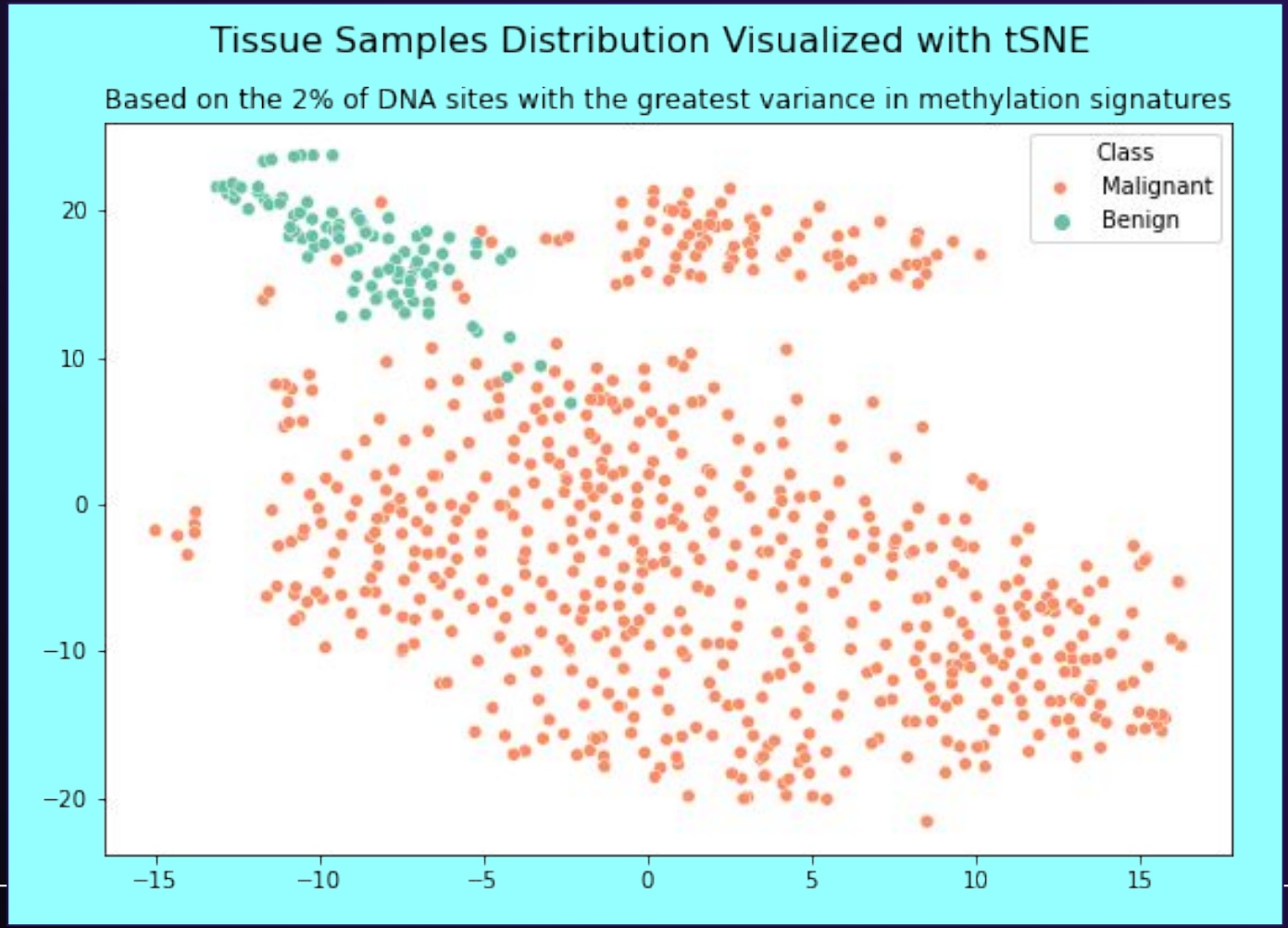


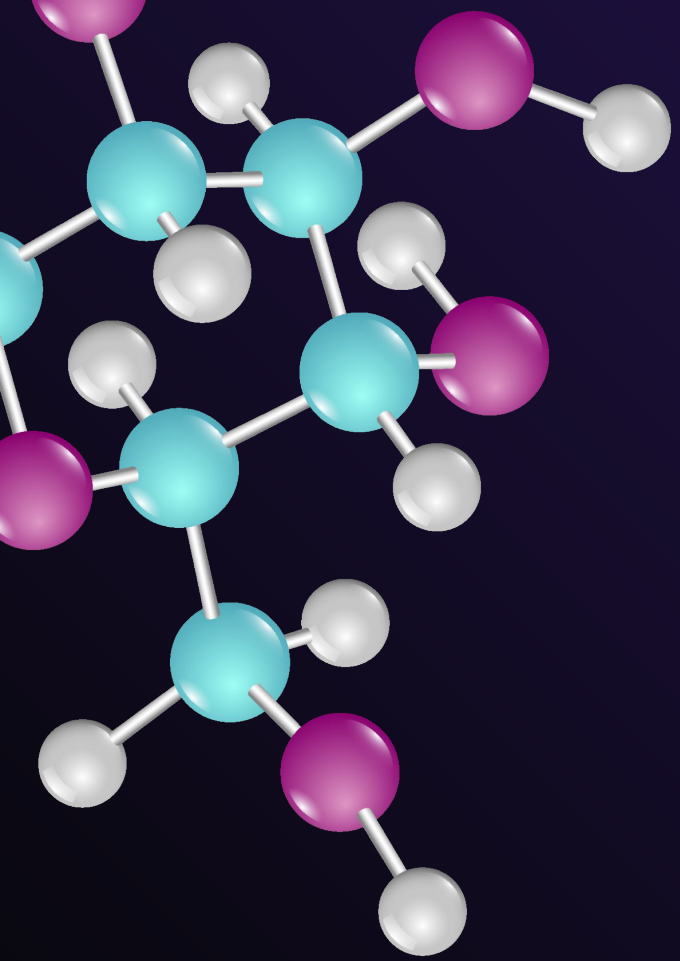


03

Unsupervised Learning
tSNE

T-distributed Stochastic Neighbor Embedding


Tissue Samples Distribution Visualized with tSNE
Based on the 2% of DNA sites with the greatest variance in methylation signatures
Class
Malignant
Benign
20
10
0
−10
−20
−15
−10
−5
0
5
10
15



04

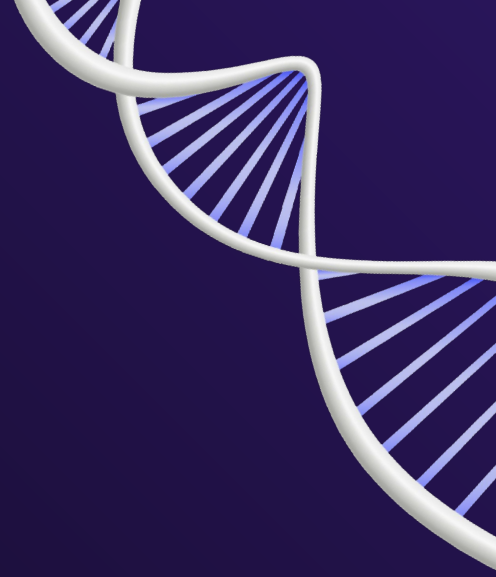
Conclusions & Recommendations

Conclusions & Recommendations

- DNA methylation signature can be utilized to detect malignant tumors
  - Researched & used in both academy and industry

- Future steps:
  - Features importance
  - Interpreting the tSNE outcomes

Questions



THANKS

DO YOU HAVE **ANY QUESTIONS?**

youremail@freepik.com
+91 620 421 838
yourwebsite.com

Please keep this slide for attribution

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon and infographics & images by Freepik