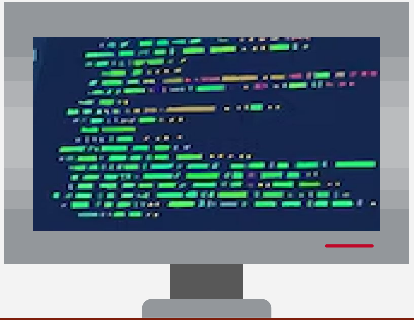
Title & Team Members



Unraveling Endometrial Cancer Heterogeneity Using Machine Learning: Somatic Mutations, Gene Expression, and ctDNA

Authors: Tomas Manea and Kiley Huffman

Table of Contents

01	**Introduction**	Explain background/context for endometrial cancer
02	**Problem Statement**	Define goals of research
03	**Methods**	Highlight key models, data sources, and techniques used
04	**Key Results**	Describe important findings
05	**Insights & Future Work**	Discuss future research, limitations, and improvements

01

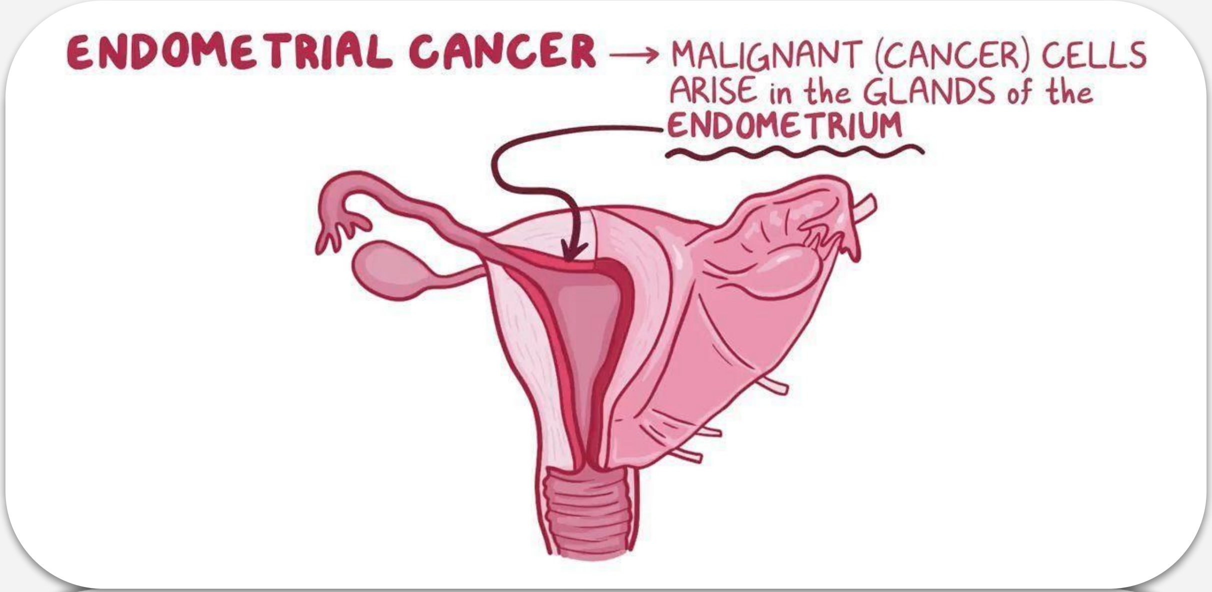
Introduction

69,120

cases reported per year

13,860

deaths projected per year


ENDOMETRIAL CANCER → MALIGNANT (CANCER) CELLS ARISE in the GLANDS of the ENDOMETRIUM

EC is a *heterogeneous* disease with multiple molecular subtypes.

It is the most common gynecologic malignancy in the U.S.

Why is studying EC important?

EC incidence has been rising steadily across all racial groups, particularly among *BIPOC* populations.

High-grade and late-stage tumors are associated with *poor prognosis* and *limited treatment options.*

Previous Studies

Molecular characterization studies have defined four primary EC subtypes with distinct genetic and prognostic profiles.

POLE ultramutated

microsatellite instability-high (MSI-H)

copy-number low

copy-number high

Integration of genomic, transcriptomic, and non-invasive biomarkers such as ctDNA remain underexplored.

02

Problem Statement

Research Question

Will a multi-omics approach integrating somatic mutation profiles, gene expression data, and circulating tumor DNA (ctDNA) provide a more comprehensive understanding of EC progression and classification?

Goals

- Uncover key molecular patterns across these data types
- Enhance subtype classification, mutation detection, and outcome prediction
- Identify strong diagnostic biomarkers and support the development of more personalized treatment strategies

03

Methods

Data Sources

1. **Somatic Mutations**: From cBioPortal, covering 197 tumors from 189 EC patients.
2. **Gene Expression**: From the GTEx portal (GTEx.gct), representing healthy endocervix tissue for comparative analysis.
3. **ctDNA Sequences**: From NCBI comprising 85 tumor-derived samples from 49 patients.

Preprocessing

Filter rare classes, drop missing values, label encode.

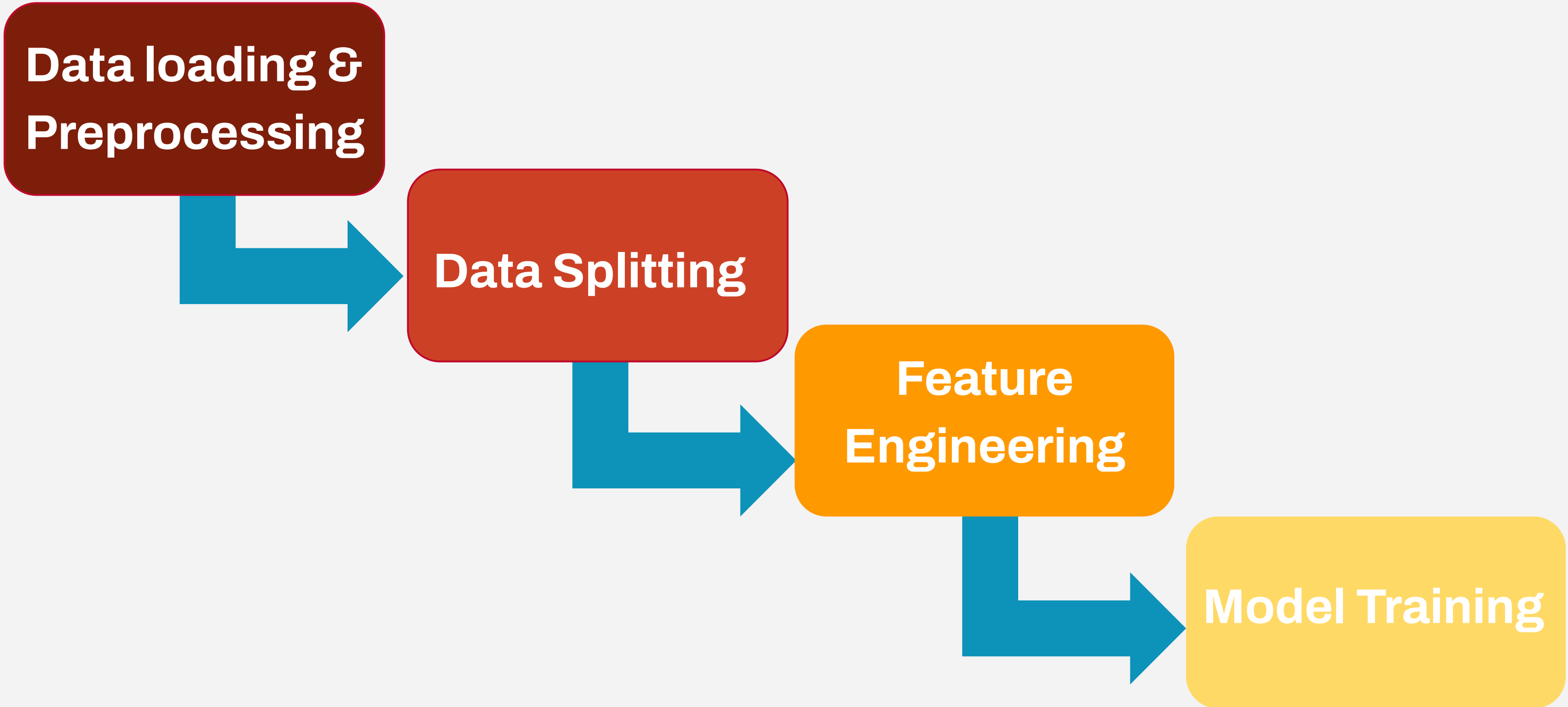
Log normalization, variance filtering, and principal component analysis (PCA) for dimensionality reduction.

Conversion to 1-mer one-hot encoding

Key Models

Task	Data	Model
Subtype Classification	Somatic mutations	Logistic Regression, SVM, Random Forest
Healthy Endocervix Profile	Gene expression	No Model
Anomaly Detection	ctDNA sequences	CNN, LSTM

Model Construction: Tumor Classification



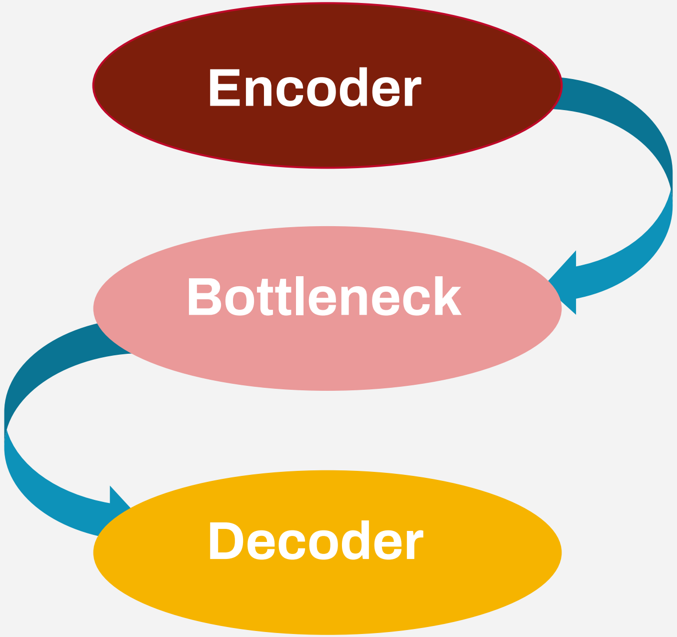
Data loading & Preprocessing

Data Splitting

Feature Engineering

Model Training

Model Construction: Anomaly Detector



04

Key Results

Table 1. AUROC/Accuracy Metrics for Cancer Subtype Classification

	AUROC (Multi-class)	Accuracy	Macro Average (Precision, Recall, F1-score)	Weighted Average (Precision, Recall, F1-score)
SVM	0.54	**0.54**	**0.30**, **0.31**, **0.26**	**0.55**, **0.54**, **0.48**
Logistic regression	**0.67**	**0.54**	**0.30**, **0.31**, **0.26**	**0.55**, **0.54**, **0.48**
Random Forest	0.50	0.36	0.19, 0.18, 0.18	0.39, 0.36, 0.37

Figure 1a.

AUROC for Cancer Subtype Classification (SVM)

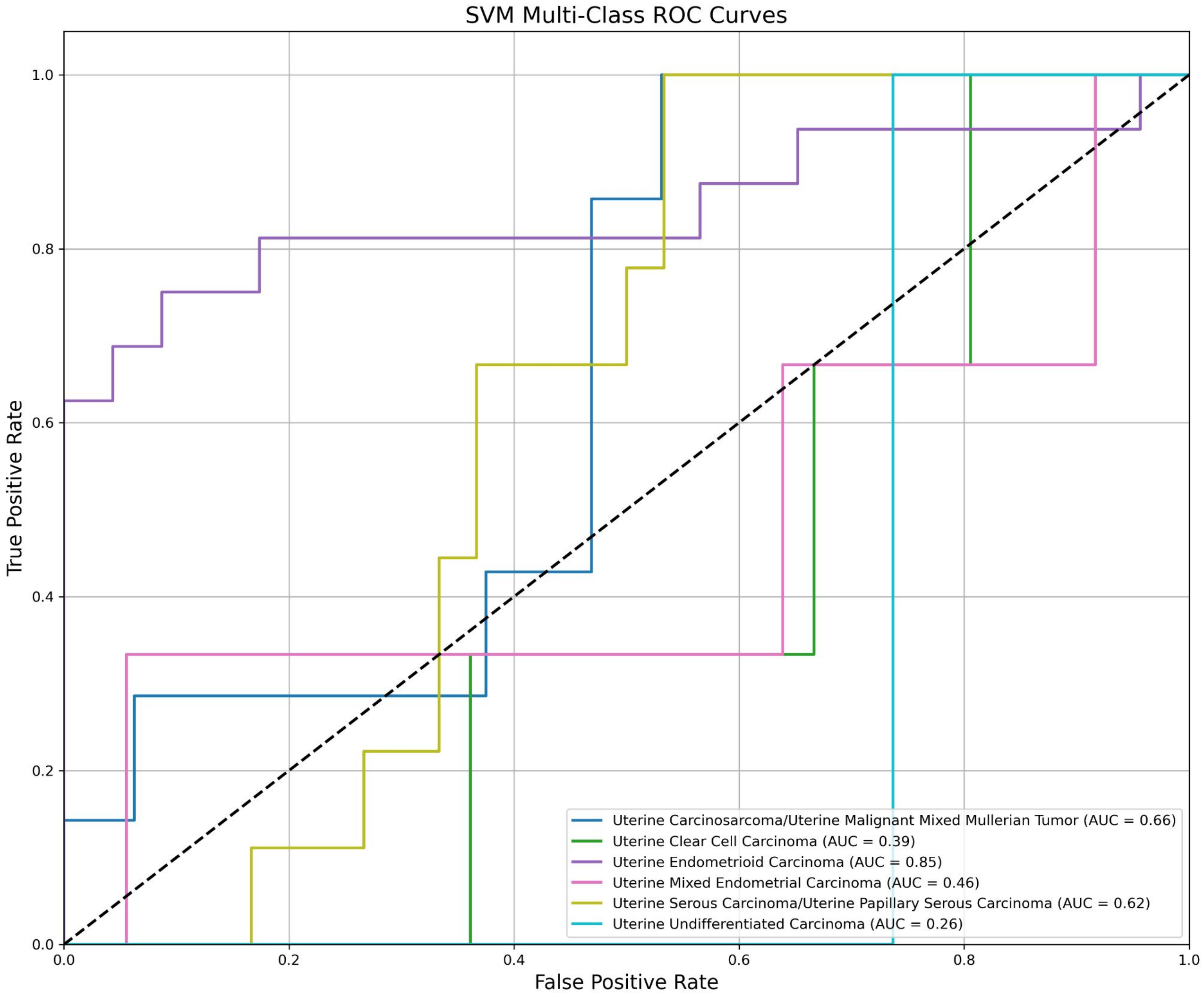


Figure 1b.

AUROC for Cancer Subtype Classification (Logistic Regression)

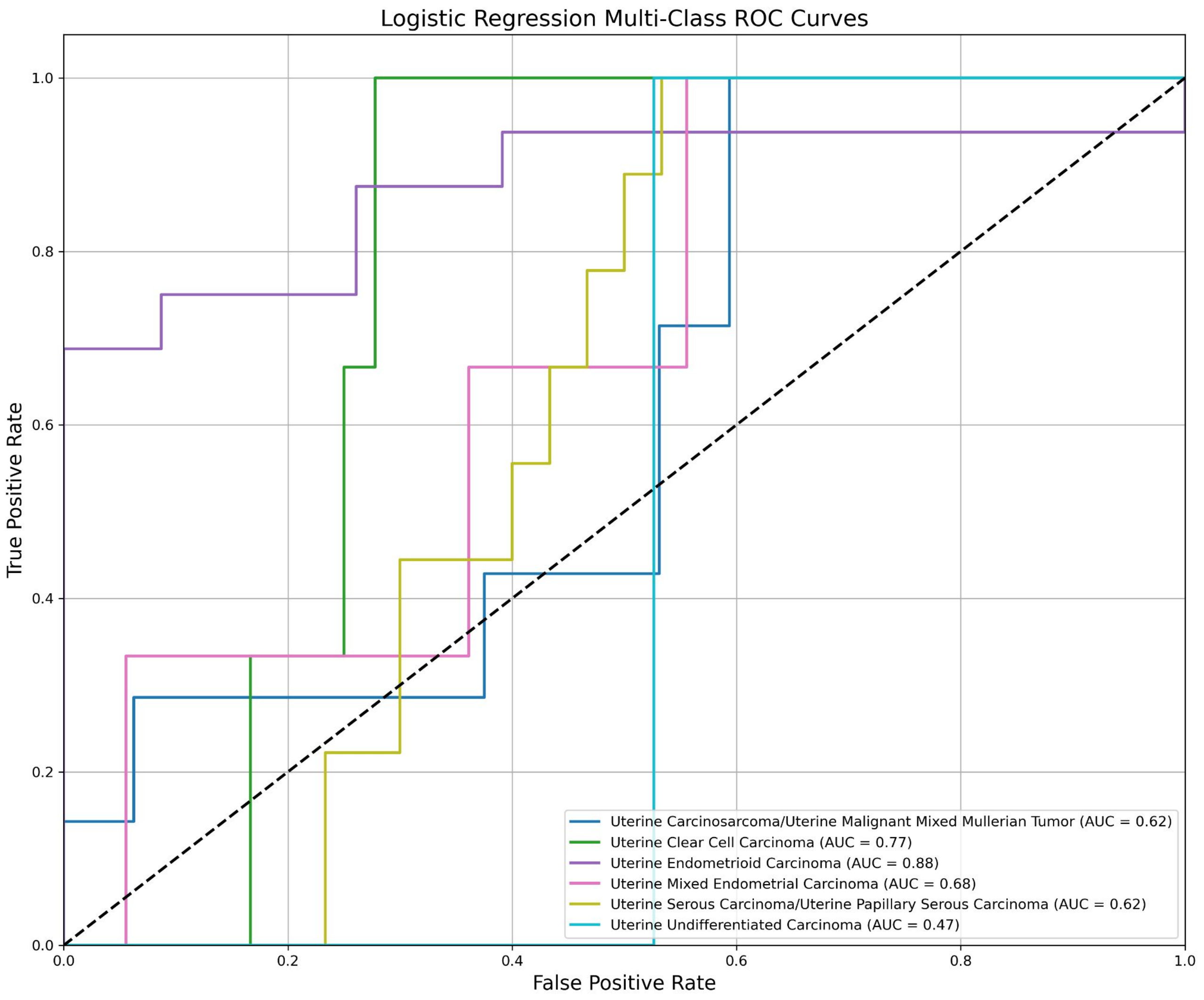
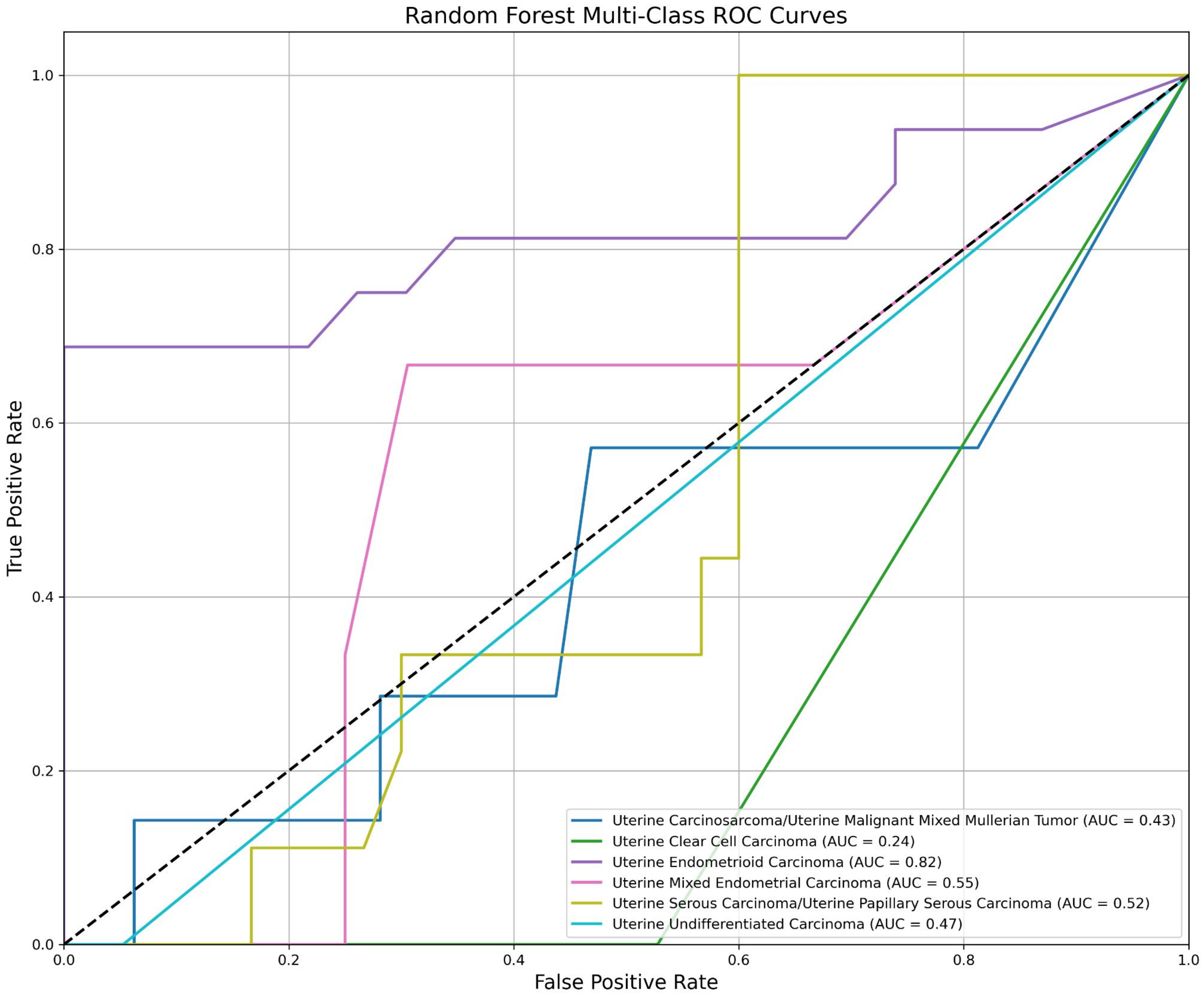


Figure 1c.

AUROC for Cancer Subtype Classification (Random Forest)



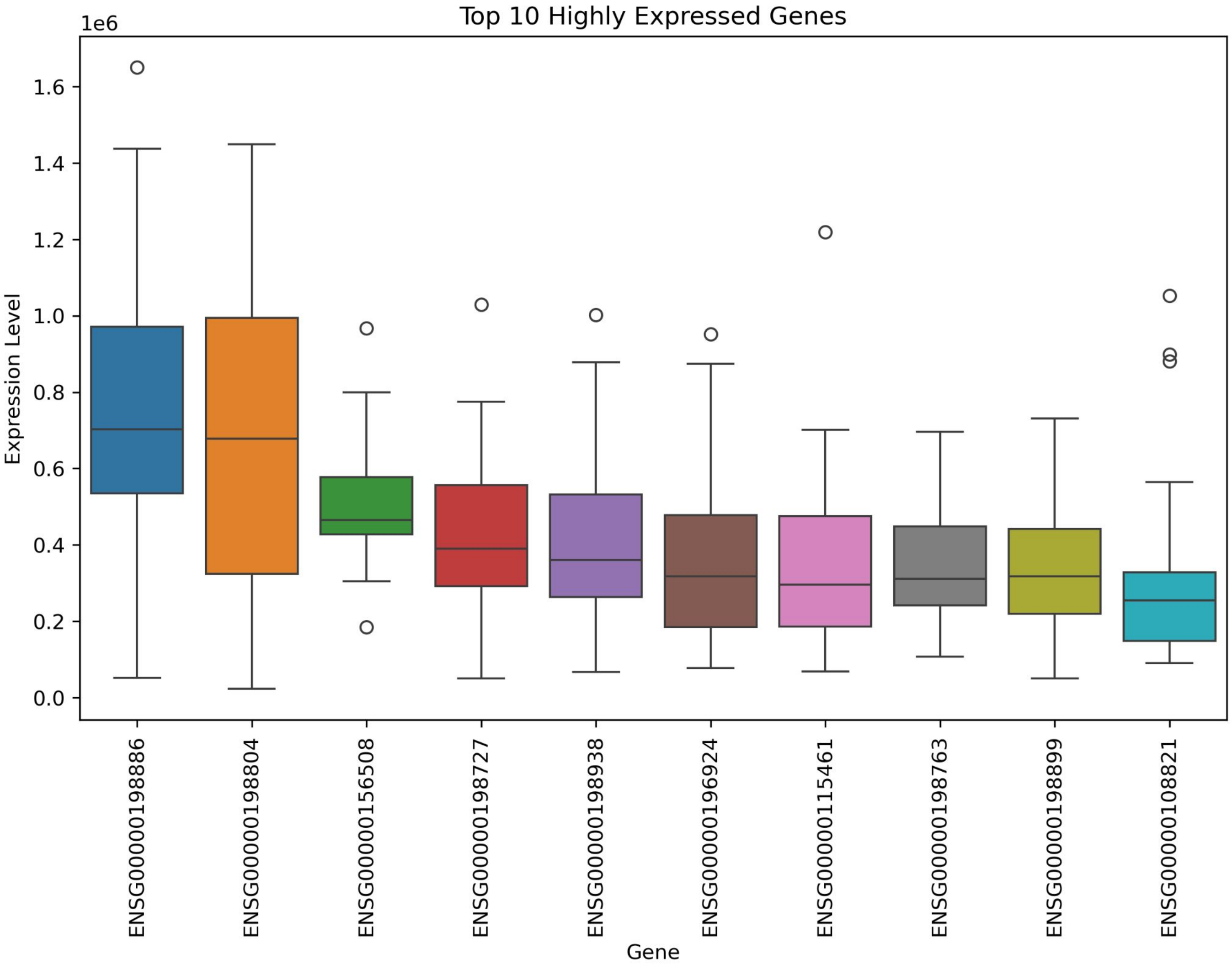


Figure 2. Highly Expressed Genes in Healthy Endocervix Tissue

Table 2. Highly Expressed Genes in Healthy Endocervix Tissue

Symbol	Name
ND4	NADH dehydrogenase subunit 4
COX1	cytochrome c oxidase subunit I
EEF1A1	eukaryotic translation elongation factor 1 alpha 1
CYTB	cytochrome b
COX3	cytochrome c oxidase subunit III
FLNA	filamin A
IGFBP5	insulin like growth factor binding protein 5
ND2	NADH dehydrogenase subunit 2
ATP6	ATP synthase F0 subunit 6
COL1A1	collagen type I alpha 1 chain

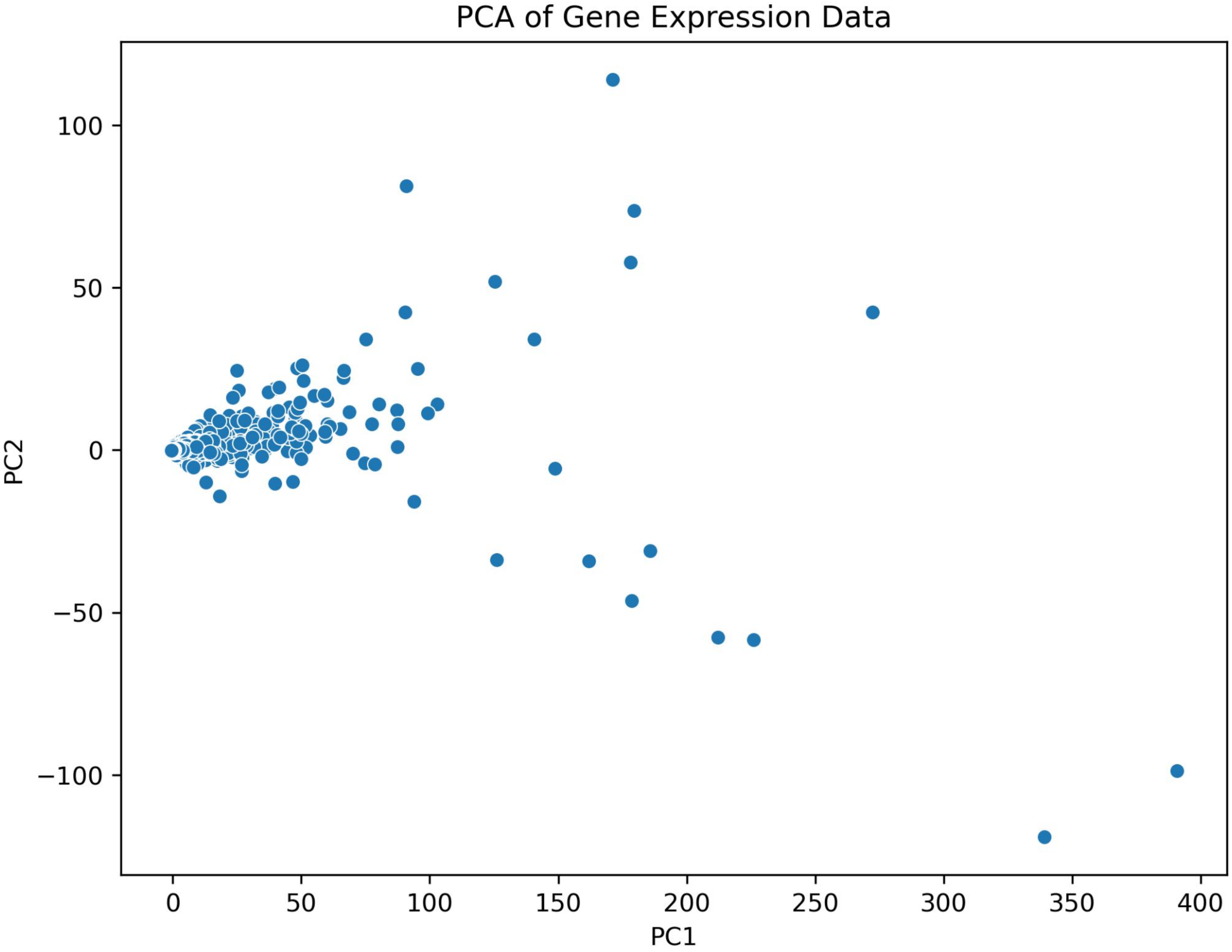


Figure 3a. PCA of Health Endocervix Tissue

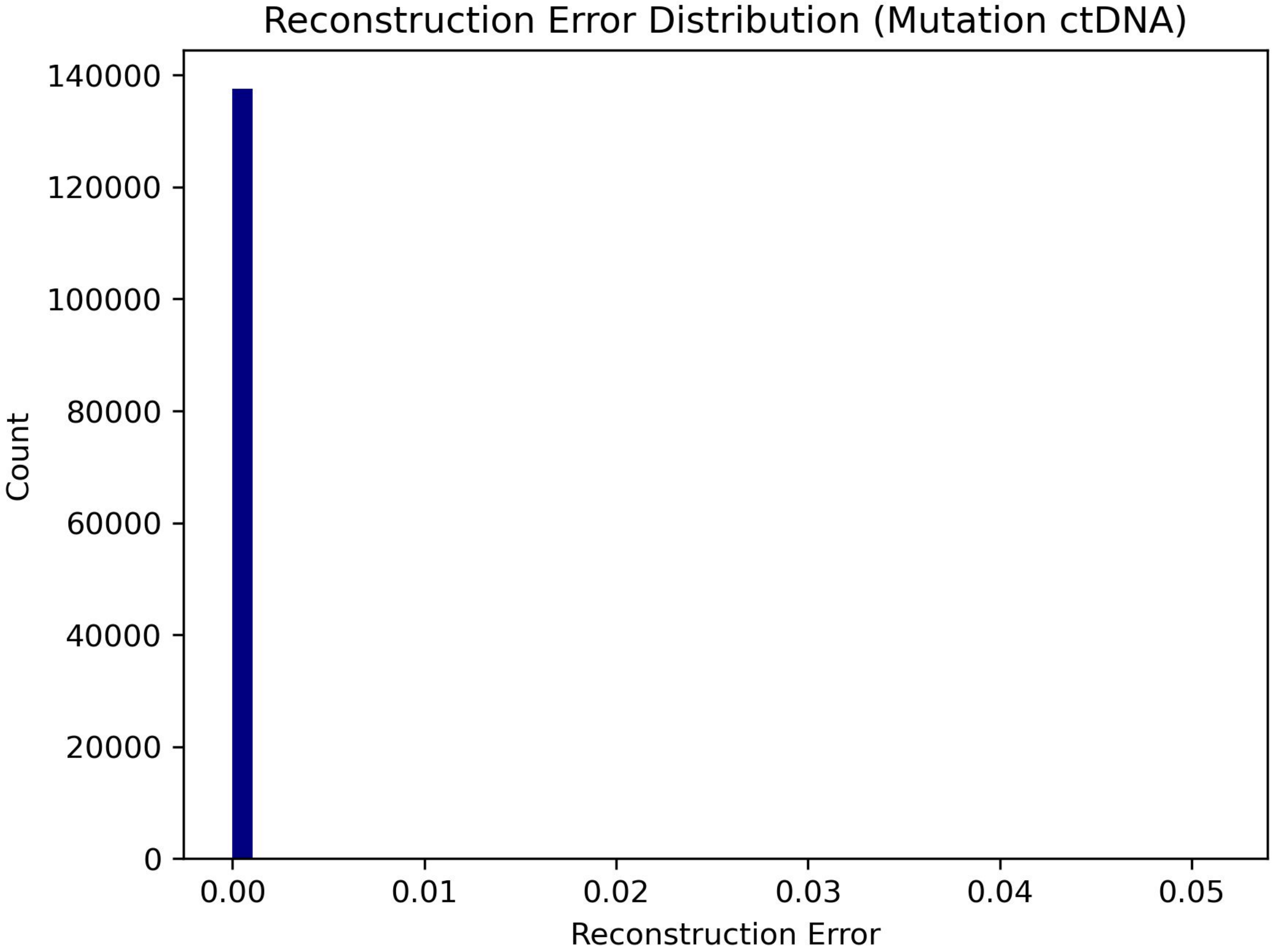


***Figure 3b.* Reconstruction Error for CNN-LSTM**

05

Insights & Future Work

Insights

- Mitochondrial genes (*ND4*, *COX1*, *CYTB*, etc.) are among the most highly expressed in *healthy* endocervix tissue
- PCA analysis did not reveal clear sample clusters; a few outliers suggest either *biological heterogeneity* or technical variation
- While our model performs well in reconstructing most ctDNA sequences, further improvements can be made to enhance its *anomaly detection* capabilities

Future Work

- Future work could involve deeper quality control to investigate the outliers and assess batch effects
- Comparative analyses against diseased or abnormal endocervical samples could help identify expression changes linked to pathology
- Functional enrichment analysis (e.g., GO terms, pathways) of the top expressed genes could yield further biological insights into endocervical tissue homeostasis and disease susceptibility.

Thanks!

Do you have any questions?

References

American Cancer Society. (2025). *Key statistics for endometrial cancer*. https://www.cancer.org/cancer/endometrial-cancer/about/key-statistics.html

Auguste, A., Genestie, C., De Bruyn, M., Alberti, N., Scoazec, J. Y., & Batteux, F. (2018). Molecular classification of endometrial carcinoma: Towards personalized treatment. *Pathology - Research and Practice*, *214*(3), 322–327. https://doi.org/10.1016/j.prp.2017.12.018

cBioPortal for Cancer Genomics. (n.d.). *MSK-IMPACT Clinical Sequencing Cohort*. https://www.cbioportal.org/

GTEx Consortium. (2015). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45*(6), 580–585. https://doi.org/10.1038/ng.2653

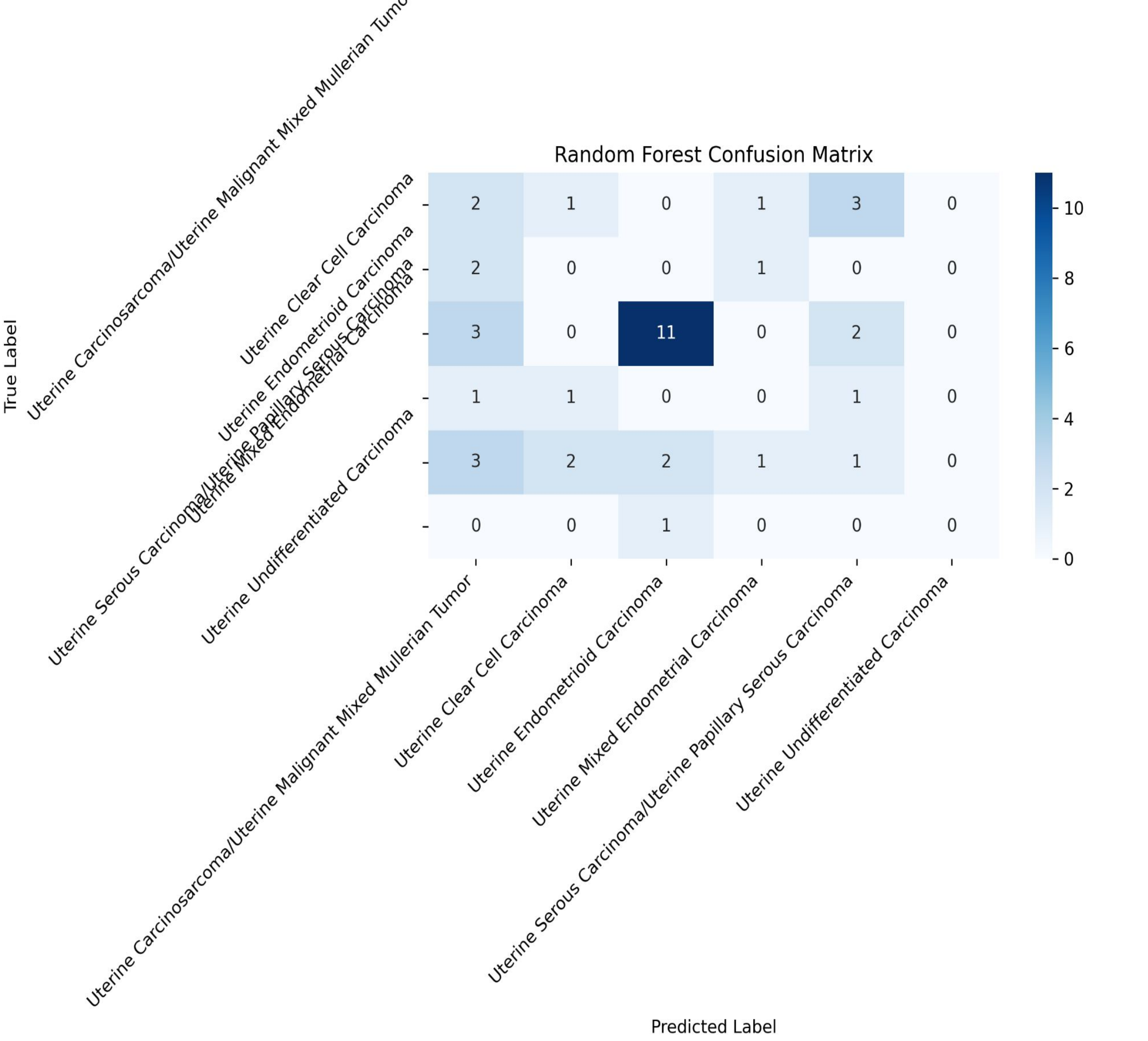
Levine, D. A., & The Cancer Genome Atlas Research Network. (2013). Integrated genomic characterization of endometrial carcinoma. *Nature*, *497*(7447), 67–73. https://doi.org/10.1038/nature12113

National Center for Biotechnology Information. (n.d.). *Sequence Read Archive: PRJDB19212 and PRJDB14089*. https://www.ncbi.nlm.nih.gov/sra

The Cancer Genome Atlas Research Network. (2013). Integrated genomic analyses of endometrial carcinoma. *Nature*, *497*(7447), 67–73. https://doi.org/10.1038/nature12113

Figure 1d.

Confusion Matrix Heatmap for Cancer Subtype Classification


Random Forest Confusion Matrix
True Label
Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumo
Uterine Clear Cell Carcinoma
Uterine Endometrioid Carcinoma
Uterine Mixed Endometrial Carcinoma
Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma
Uterine Undifferentiated Carcinoma
Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor
Uterine Clear Cell Carcinoma
Uterine Endometrioid Carcinoma
Uterine Mixed Endometrial Carcinoma
Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma
Uterine Undifferentiated Carcinoma
Predicted Label