Tomas Manea and Kiley Huffman
QBIO 465: AI in Biology and Medicine
Final Project Proposal
Spring 2025

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

Abstract

Endometrial cancer (EC) is a heterogeneous disease with multiple molecular subtypes that influence prognosis and treatment response. A multi-omics approach integrating somatic mutation profiles, gene expression data, and circulating tumor DNA (ctDNA) has the potential to provide a more comprehensive understanding of EC progression and classification. This study aims to uncover key molecular patterns across these data types using machine learning techniques including logistic regression, support vector machines (SVMs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory networks (LSTMs). We will also employ ensemble and survival models to enhance subtype classification, mutation detection, and outcome prediction. By integrating these modalities, we hope to identify strong diagnostic biomarkers and support the development of more personalized treatment strategies.

Introduction

Endometrial cancer is the most common gynecologic malignancy in the United States, with approximately 69,120 new cases and 13,860 deaths projected each year (American Cancer Society, 2025). While many cases are diagnosed at an early stage with favorable outcomes, high-grade and late-stage tumors are associated with poor prognosis and limited treatment options. EC incidence has been rising steadily across all racial groups, particularly among non-white populations. Molecular characterization studies, including one done by The Cancer Genome Atlas (TCGA), have defined four primary EC subtypes—POLE-ultramutated, microsatellite instability-high (MSI-H), copy-number low, and copy-number high—with distinct genetic and prognostic profiles (Levine et al., 2013). Other classification schemes propose up to five subtypes with implications for therapy and recurrence risk (Auguste et al., 2018). Despite these advances, the integration of genomic, transcriptomic, and non-invasive biomarkers such as ctDNA remains underexplored. This study proposes a machine learning pipeline to classify EC subtypes, detect mutations from ctDNA, and predict survival outcomes using a multi-omic dataset. Our integrative approach seeks to bridge the gap between molecular complexity and clinical application in EC diagnostics and treatment.

Methodology

Data Acquisition and Preprocessing

We will integrate and process three primary datasets:

1. Somatic Mutations: From cBioPortal (MSK-IMPACT.tsv), covering 197 tumors from 189 EC patients. Preprocessed using one-hot encoding of high-frequency driver mutations (e.g., TP53, PIK3CA, PTEN, ARID1A). Additional derived features may include mutation burden and pathway-level scores.
2. Gene Expression: From the GTEx portal (GTEx.gct), representing healthy endocervix tissue for comparative analysis. Preprocessed using log normalization, variance filtering, and principal component analysis (PCA) for dimensionality reduction.
3. ctDNA Sequences: Two FASTA files from NCBI (PRJDB19212 and PRJDB14089), comprising 85 tumor-derived samples from 49 patients. Preprocessed using conversion to k-mer frequency vectors (k = 6 or 7), with potential use of embeddings for sequence modeling.

We will ensure proper normalization, batch correction (if needed), and patient-level partitioning to prevent data leakage.

Model Building

We will build several models using machine learning techniques in Python. Model tuning will involve cross-validation and hyperparameter optimization (e.g., grid search, random search). For deep learning models, we will monitor overfitting using dropout, early stopping, and data augmentation where applicable.

Task	**Data**	**Model**
EC Subtype Classification	Somatic mutations	Logistic Regression, SVM
Tumor subtyping	Gene expression	CNN
Mutation Detection	ctDNA sequences	RNN, LSTM
Multi-Omics Integration	Combined features	Random forest, multi-input neural networks
Survival Prediction	All omics, metadata	Cox proportional hazards, gradient boosting survival trees

Evaluation and Visualization

We will evaluate performance with task-specific metrics; for subtype classification, we will use accuracy, F1-score, and AUROC curves; for ctDNA mutation detection, we will use precision-recall curve, sensitivity, and specificity; for multi-omics integration, we will use feature importance (SHAP values) and a confusion matrix; for survival modeling, we will use concordance index, Kaplan-Meier plots, and log-rank tests. Model results will be visualized using libraries such as matplotlib, seaborn, and lifelines.

Expected Results

We expect that multi-omic integration will significantly improve EC subtype classification compared to single-modality models. We predict that the CNNs will outperform linear models in gene expression-based classification due to their ability to capture spatial gene relationships, and that the LSTM models will show strong performance in mutation detection from ctDNA, potentially supporting the development of non-invasive diagnostics. It is expected that survival prediction models will successfully stratify patients into high- and low-risk groups based on molecular features, and that key biomarkers identified from feature attribution techniques (e.g., SHAP) may yield novel diagnostic or therapeutic targets.

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