Tomas Manea and Kiley Huffman

QBIO 465: AI in Biology and Medicine
Final Project Report
May 6, 2025

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

Abstract

To improve subtype classification and molecular diagnostics of endometrial cancer, we applied an integrative approach combining supervised machine learning, gene expression analysis, and deep learning. Using genomic features from the MSK-IMPACT dataset, we evaluated three classification models—Logistic Regression, Support Vector Machine, and Random Forest—for subtype prediction. Logistic Regression achieved the best performance (AUROC = 0.67), while Random Forest underperformed, highlighting the limitations of complex models under data imbalance and feature constraints. Gene expression analysis of healthy cervix-endocervix tissue identified strong expression of mitochondrial (e.g., ND4, COX1, CYTB) and structural (e.g., FLNA, COL1A1) genes, consistent with the tissue's high metabolic activity. Principal Component Analysis (PCA) revealed limited heterogeneity and a small number of outliers, suggesting potential batch effects or biological variation. Additionally, a CNN-based autoencoder trained on circulating tumor DNA (ctDNA) sequences effectively reconstructed most input sequences and flagged potential anomalies through elevated reconstruction errors. These findings highlight the diagnostic potential of integrative computational models while outlining key challenges, such as class imbalance and limited feature discriminability, in subtype classification. Future work should explore multimodal data fusion and model refinement to enhance classification accuracy and biological insight.

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, late-stage tumors have limited treatment options. The incidence of endometrial cancer has been rising steadily, particularly among BIPOC populations. Molecular characterization studies 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. Our study seeks to answer the following 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?

Our main objectives are to (1) uncover key molecular patterns to enhance subtype classification, mutation detection, and outcome prediction and (2) to identify strong diagnostic biomarkers to support the development of more personalized treatments.

Methods

Data Acquisition and Preprocessing

We utilized three datasets:

1. Somatic Mutations: From cBioPortal (MSK-IMPACT.tsv), covering 197 tumors from 189 patients. Preprocessed using one-hot encoding of high-frequency driver mutations (e.g., TP53, PIK3CA, PTEN, ARID1A).
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.

Model Building

Table A. Models, data used, and their tasks.

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

Supervised Learning on MSK-IMPACT Clinical Data

We curated a subset of the MSK-IMPACT clinical dataset, focusing on four features: Mutation Count, Tumor Mutational Burden (TMB; nonsynonymous only), Fraction Genome Altered, and Neoplasm Histologic Grade. To ensure statistical validity, only cancer types with at least two samples were retained. After removing entries with missing values, the target variable (Cancer Type Detailed) was label-encoded into integers for model compatibility. All numerical input features were standardized using StandardScaler to have zero mean and unit variance, enabling fair comparison across models. The dataset was split into 80% training and 20% testing subsets using stratified sampling to preserve class distributions. Three supervised models were trained on the standardized feature set:

1. **Support Vector Machine (SVM)**—RBF kernel and one-vs-rest (OvR) decision function
2. **Logistic Regression (LR)**—OvR classification and a maximum of 1000 iterations
3. **Random Forest (RF)**—Classifier with 100 decision trees

The models were evaluated using accuracy (proportion of correctly classified samples), AUROC (one-vs-rest multiclass evaluation using predicted probabilities), and classification reports (precision, recall, and F1-score per class).

CNN-LSTM-Based Autoencoder for ctDNA Sequences

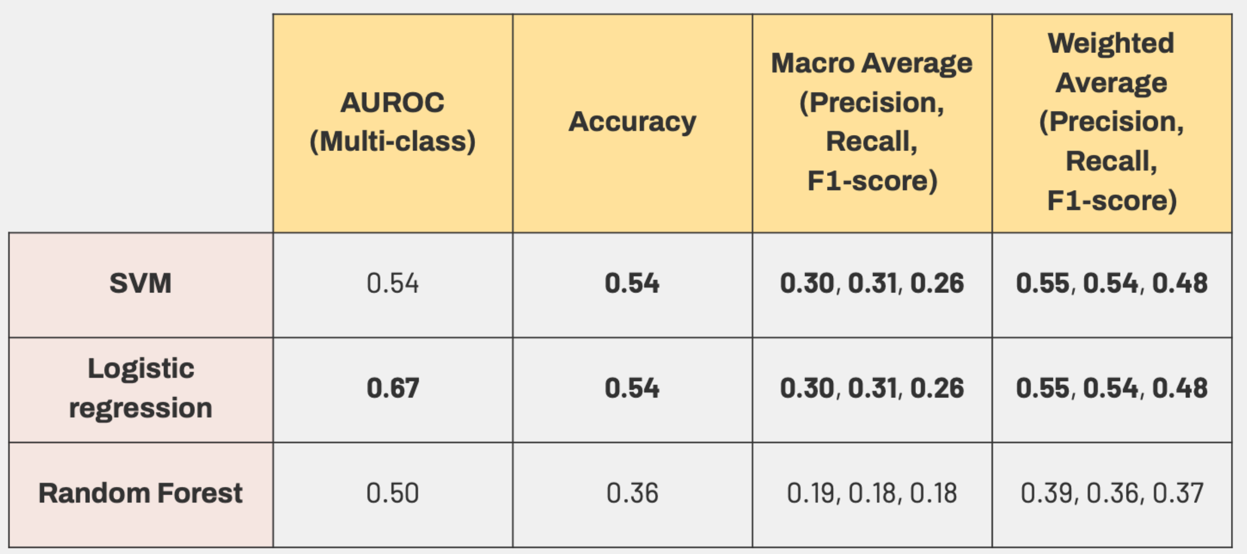
We also developed a CNN-based autoencoder to analyze cell-free DNA (ctDNA) sequences. DNA sequences were parsed from FASTA files using Biopython's SeqIO, extracted as strings, and one-hot encoded, mapping nucleotides (A, C, G, T) to binary vectors. To ensure uniformity, all sequences were truncated or zero-padded to a fixed length of 25 nucleotides. The autoencoder architecture included an Encoder with Conv1D (32 filters, kernel size 3, ReLU activation) and MaxPooling1D (pool size 2), a Bottleneck with Conv1D (16 filters, kernel size 3, ReLU activation), and a Decoder with UpSampling1D (doubles sequence length), Conv1D (4 filters, kernel size 3, sigmoid activation), and Cropping1D (adjusts output length). The model was compiled with the Adam optimizer and trained to minimize mean squared error (MSE). Data was split randomly into 80% training and 20% testing subsets. Training was conducted with a batch size of 128 over 10 epochs. Post-training, reconstruction error (MSE between original and reconstructed sequences) was used to identify anomalous patterns. A histogram of reconstruction errors was plotted to visualize error distributions, enabling detection of potentially mutated vs. healthy ctDNA sequences.

The code for our study can be found on GitHub at the link below:

https://github.com/TDManEA/QBIO465/blob/main/qbio465_final_project.ipynb

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

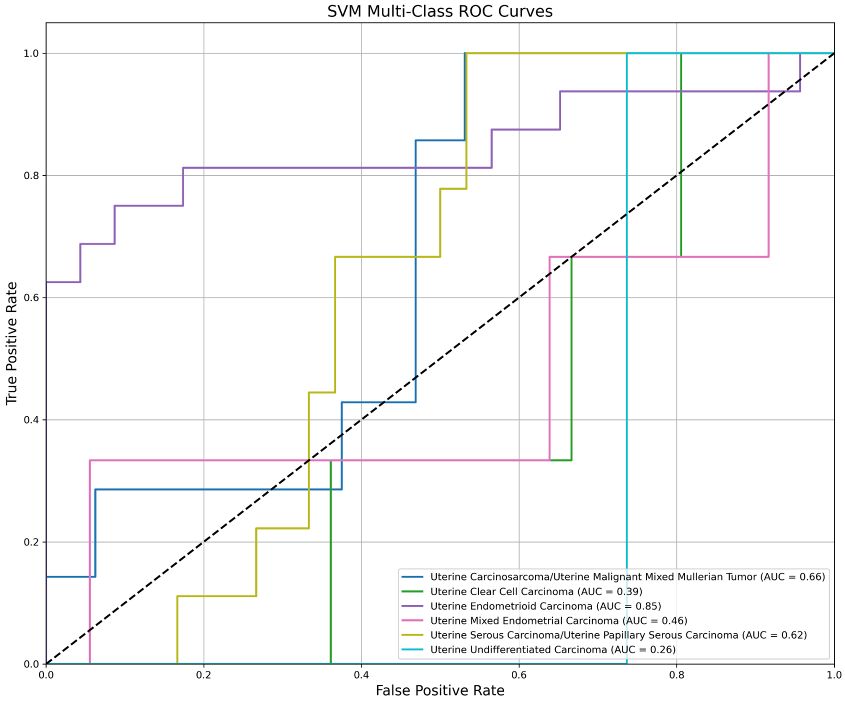


Figure 1a. Support Vector Machine (SVM): AUROC for Cancer Subtype Classification

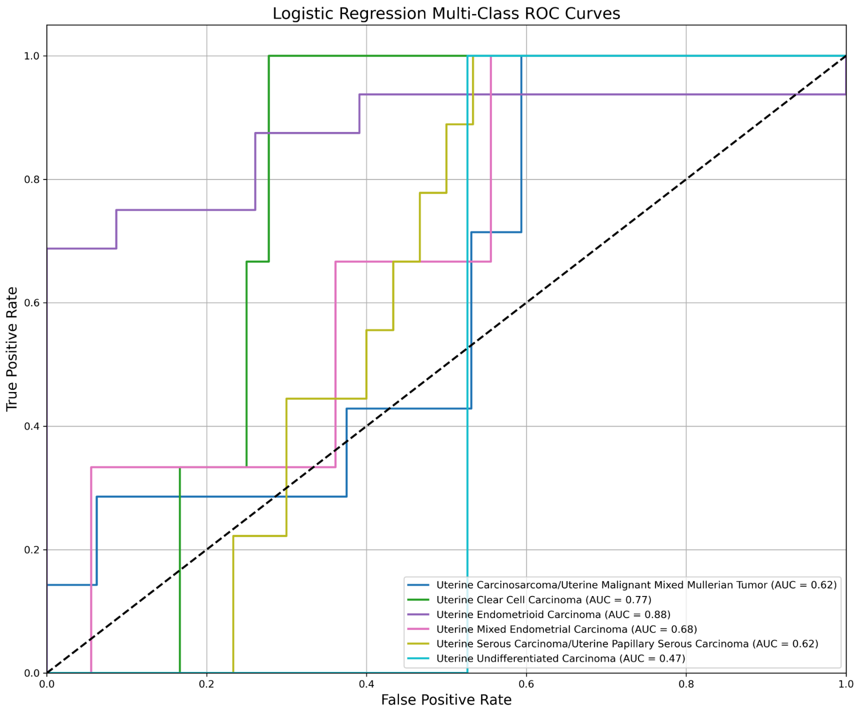


Figure 1b. Logistic Regression (LR): AUROC for Cancer Subtype Classification

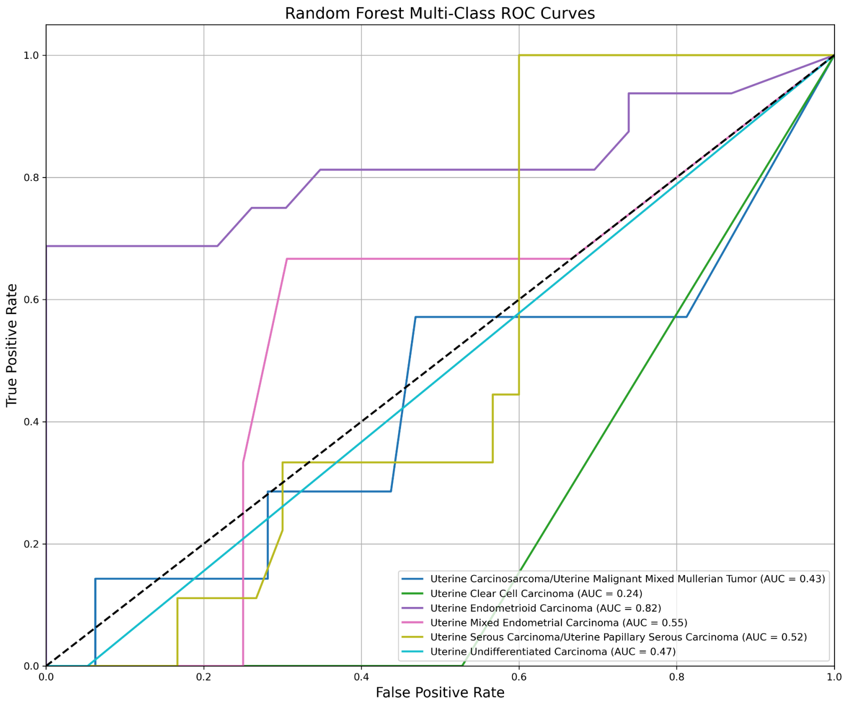


Figure 1c. Random Forest: AUROC for Cancer Subtype Classification

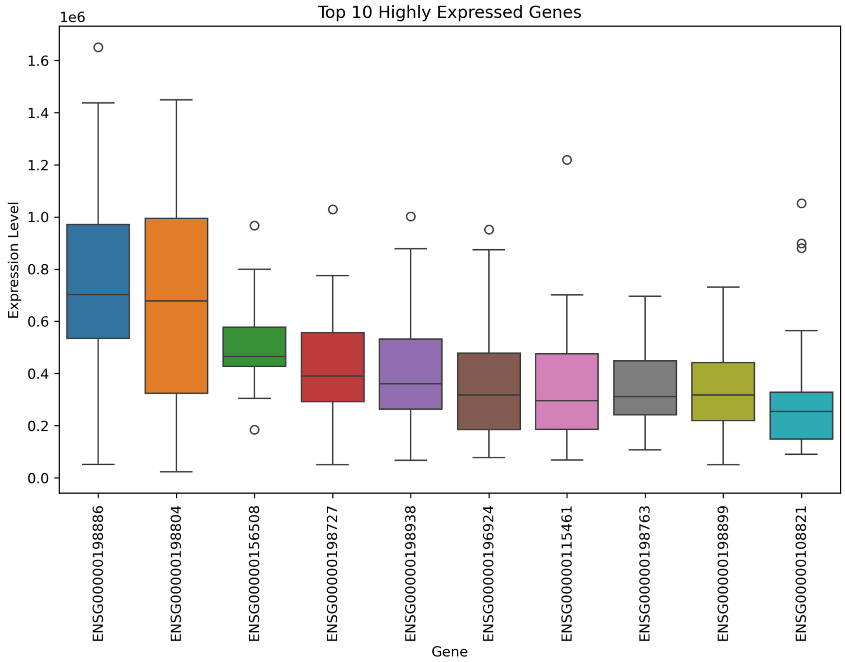


Figure 2. Highly expressed genes in healthy endocervix tissue.

Table 2. Symbols and names of the 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

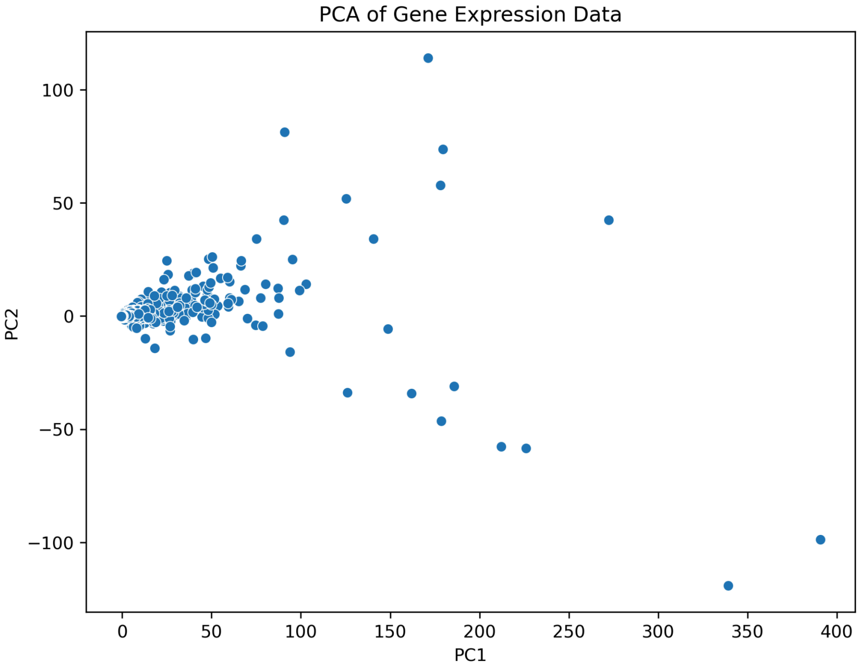


Figure 3. Principal Component Analysis (PCA) of healthy endocervix tissue.

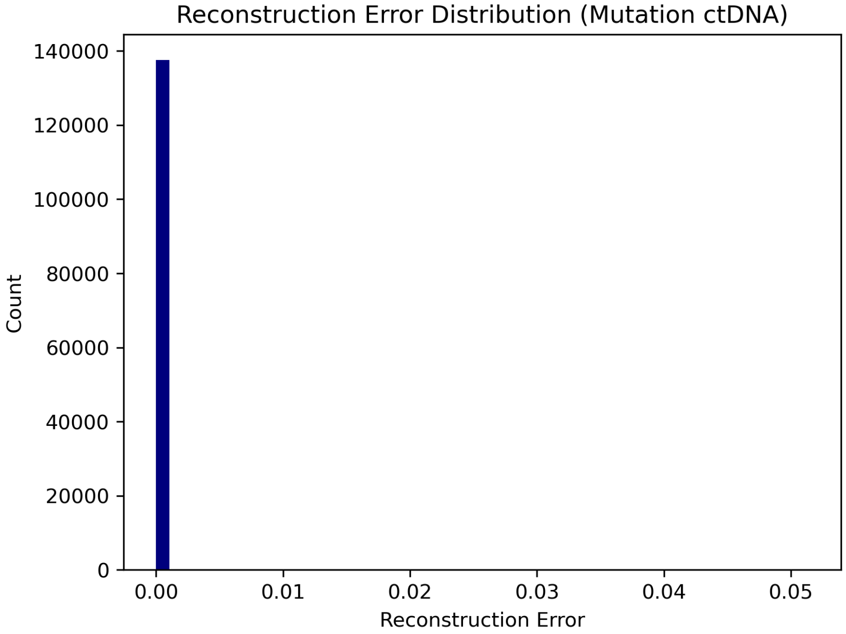


Figure 4. Reconstruction Error for CNN-LSTM.

Discussion

Class Prediction

Among the evaluated classification algorithms, Logistic Regression demonstrated the best overall performance, achieving the highest area under the receiver operating characteristic curve (AUROC = 0.67) and matching the support vector machine in accuracy (54%), as shown in ***Table 1***, ***Figure 1a***, and ***Figure 1b***. In contrast, the Random Forest model underperformed, with an AUROC of 0.50 and accuracy of 36%, as shown in ***Table 1*** and ***Figure 1c***. These results suggest that simpler linear models, such as Logistic Regression, may outperform more complex, nonlinear approaches for cancer subtype classification. The limitations of the Random Forest model, including potential overfitting or insufficient training data, may account for this discrepancy. Our model's performance is modest compared to existing studies that utilize more advanced methods. For instance, Kim et al. (2021) achieved an AUROC of 0.944 for classifying endometrioid and serous endometrial adenocarcinomas using convolutional neural networks (CNNs) applied to whole-slide images. Similarly, the Panoptes2 model (2020), a multi-resolution deep learning approach, reported an AUROC of 0.969 for per-patient classification of endometrial cancer histological subtypes.

Despite the moderate performance of our three models, class-specific outcomes varied considerably. All models achieved high precision and recall for Uterine Endometrioid Carcinoma, indicating that this subtype possesses distinct gene expression patterns that facilitate accurate classification. Conversely, Uterine Serous Carcinoma exhibited moderate recall but poor precision, suggesting frequent false positives. This may reflect the molecular overlap between serous endometrial and high-grade serous ovarian cancers, which can complicate subtype differentiation based on gene expression alone (Kommoss et al., 2018). Our models struggled most with rare subtypes, such as Clear Cell Carcinoma, Mixed Endometrial Carcinoma, and Undifferentiated Carcinoma, with precision and recall scores often approaching zero. This highlights the impact of class imbalance, where the underrepresentation of rarer classes hindered the model's ability to accurately learn these patterns. Additionally, the low AUROC values across all models (≤ 0.67) suggest limited discriminative power of the available gene expression features. To mitigate this, we would need to focus on more robust feature engineering or the inclusion of complementary data.

Gene Expression in Cervix-Endocervix Tissue

As shown in ***Table 2*** and ***Figure 2***, several genes were highly expressed in healthy cervix-endocervix tissue, particularly those involved in mitochondrial function (e.g., ND4, COX1, CYTB, COX3), structural integrity (e.g., FLNA, COL1A1), and protein synthesis (e.g., EEF1A1, IGFBP5). The prominence of mitochondrial genes suggests a high metabolic demand characteristic of this tissue, consistent with prior studies highlighting metabolic gene upregulation across endometrial subtypes (Lähdesmäki et al., 2021). Principal Component Analysis (PCA) revealed that the first two components captured most of the variance in gene expression, as shown in ***Figure 3***. However, PCA did not uncover distinct clusters, suggesting limited tissue heterogeneity or the influence of unmeasured confounders. A small number of outliers were observed, potentially indicating biological heterogeneity, technical variation, or batch effects. While comparisons with studies focused on this specific tissue type are limited, our findings align with broader research indicating that mitochondrial genes play a significant role in cellular metabolism and function across different tissues.

Autoencoder Reconstruction of ctDNA Sequences

A CNN-based autoencoder was trained to reconstruct circulating tumor DNA (ctDNA) sequences. The model achieved a significant reduction in reconstruction loss during training, with a final validation loss of approximately 5.35e-06 (***Figure 4***), indicating strong learning of typical sequence patterns. The left-skewed distribution of reconstruction errors for the test set suggests that the model successfully reconstructed most ctDNA sequences. However, sequences with high reconstruction errors may represent genomic anomalies or mutations, a pattern also observed in prior deep learning studies using autoencoders for somatic variant detection in sequencing data (Zhou et al., 2019). These findings suggest that further refinement of the autoencoder could improve its capacity for anomaly detection and ctDNA-based diagnostics.

Future Directions

Our study faced several limitations, primarily class imbalance, which hindered the performance of models on rare subtypes. Future work should incorporate techniques such as synthetic oversampling or class-weighted loss functions to address this issue. Additionally, incorporating additional data (e.g., clinical, histopathological, or epigenomic) could improve subtype discriminability and predictive power. Investigating differential gene expression between healthy and diseased endocervix samples could offer insights into molecular mechanisms driving subtype-specific signatures. If we were to revisit this study, we would further refine the autoencoder architecture to enhance its utility for mutation detection and ctDNA-based diagnostics.

While the performance metrics in our study are modest compared to studies utilizing deep learning models, our findings provide valuable insights into the challenges of classifying uterine cancer subtypes using gene expression data. The difficulties with rare subtypes underscore the need for improved model architectures and strategies to address class imbalance. For example, if we were to revisit this study, we would further refine the autoencoder architecture to enhance its utility for mutation detection and ctDNA-based diagnostics. Future research should explore the integration of multi-modal data and advanced modeling techniques to enhance classification accuracy and provide deeper biological insights into endometrial cancer pathogenesis.

Member Contribution (Percentage)

We divided all of the work equally;
Tomas Manea: 50%
Kiley Huffman: 50%

Appendix: Code Repository

The code for this study can be found on GitHub at this link:
https://github.com/TDManEA/QBIO465/blob/main/qbio465_final_project.ipynb

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

Kim, J., Lee, J., Kim, K., Kim, J., & Lee, C. (2021). *Deep learning-based classification of endometrial adenocarcinoma subtypes using whole-slide images*. *Journal of Pathology Informatics, 12*(1), 1-8. https://doi.org/10.4103/jpi.jpi_20_21

Kommoss, S., McConechy, M. K., Kommoss, F., Leung, S., Bunz, A., Magrill, J., ... & Huntsman, D. G. (2018). Molecular subtypes of high-grade endometrial carcinomas: New insights from integrated genomic analysis. *The Journal of Pathology, 245*(4), 443–452. https://doi.org/10.1002/path.5091

Lähdesmäki, H., Pirskanen, A., & Hautaniemi, S. (2021). Gene expression variation in endometrial carcinoma subtypes. *Nature Communications, 12*, 3452. https://doi.org/10.1038/s41467-021-23867-y

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

Panoptes2. (2020). *Panoptes2: A multi-resolution deep learning model for per-patient classification of endometrial cancer histological subtypes*. Nature Communications, *11*(1), 1184. https://doi.org/10.1038/s41467-020-15023-z

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

Zhou, X., Liu, C., Xu, J., & Zhang, M. (2019). Deep learning for mutation detection in next-generation sequencing data. *Nature Biomedical Engineering, 3*(3), 190–199. https://doi.org/10.1038/s41551-018-0333-5