Original Research Article

The Biological Significance of cfDNA Methylation Patterns in Early Ovarian Cancer Detection and Analytical Methods for Detection

Blossom Patel[1] and Hangpeng Li[2]

[1]*South Brunswick High School, 750 Ridge Rd, Monmouth Junction, NJ 08852, United States;*
[2]*University of Oxford, Oxford, OX1 2JD, United Kingdom*

ABSTRACT

Ovarian cancer is one of the deadliest gynecological cancers; survival over five years drops sharply, from about 90% when found early to under 30% if detected late. Although ovarian cancer detection based on blood tests that analyze cell-free DNA (cfDNA) could provide a less invasive option, finding consistent signals in the limited data found in blood samples is difficult. To address the high-dimensional nature of this data, the critical challenge of data leakage in machine learning pipelines was investigated. Taken together, combining sophisticated AI methods with highly sensitive methylation tests appears to offer the best chance for developing practical early-detection screening tools. A synthetic dataset was generated that mirrors plasma cfDNA methylation fragments, and a simulation was performed to determine the number of cancerous versus noncancerous differentially methylated regions (DMRs). First, a standard "global" selection model was simulated, which displayed the artificial inflation (overfitting) of accuracy that occurs due to data leakage. Second, a nested-CV elastic-net model was tested to isolate the true biological signal from cfDNA methylation. After modeling this procedure, the leakage-safe model could successfully distinguish early ovarian cancer from non-tumor samples (AUROC 0.938; AUPRC 0.914). This project lays the foundation for future exploration of theory-driven, AI-powered liquid biopsy models.

Keywords: Ovarian Cancer (OC); cell-free circulating DNA (cfDNA) methylation; Liquid Biopsy (LB) biomarkers; machine learning (ML); differentially methylated regions (DMRs)

INTRODUCTION

The 5-year relative survival rate for ovarian cancer is 51.6%, making ovarian cancer (OC) the most deadly gynecological cancer (1). About 3 quarters of OC patients are diagnosed at a late stage, due to the asymptomatic nature of early-stage disease, which often precludes complete surgical resection. Early detection not only increases the 5-year relative survival rate but also leads to surgery that is less complicated (2).

Cell-free DNA (cfDNA) offers a promising pathway for early cancer detection (3); it consists of short DNA fragments that are released from all types of cells in the body (4). Within this pool, circulating tumor DNA (ctDNA) harbors specific epigenetic changes, such as silencing of tumor suppressor genes. These alterations serve as valid indicators of malignancy. Differentially methylated regions (DMRs) are short stretches of DNA (often hundreds of bases) in which methylation levels differ consistently between two groups. Because these DMRs occur only in tumors and appear early in disease

Corresponding author: Blossom Patel, E-mail: blossompatel2677@gmail.com.

Accepted May 4, 2026
https://doi.org/10.70251/HYJR2348.431619

progression, they are an ideal biomarker for early detection of ovarian cancer (4).

However, the clinical utility of ctDNA is often limited by its relative scarcity compared with total cfDNA. CtDNA represents only a small fraction of the total pool, so pinpointing a single mutation can be technically difficult. To combat this, a methylation-based assay can simultaneously scan for thousands of tumor-specific methylation marks, rather than pinpointing rare individual mutations, thereby creating a robust, detectable signal (3). To preserve this signal, non-destructive analytical methods are required. Cell-free Methylated DNA Immunoprecipitation and high-throughput Sequencing (cfMeDIP-seq) is an advantageous method because it avoids the destructive chemical degradation of traditional approaches (2).

To handle the extensive genetic data that is produced by cfMeDIP-seq, robust machine learning is required. However, a critical knowledge gap exists in how these computational models process sparse cfDNA data. Because the number of genomic features far exceeds the limited number of patient samples available, machine learning pipelines frequently overfit the data. The objective of this study is to establish a methodologically sound computational pipeline for early detection of ovarian cancer. It is hypothesized that by restricting feature selection strictly within training folds, a machine learning model can accurately identify true DMRs without artificial inflation.

The significance of DMRs in cfDNA methylation was analyzed by running simulations to determine how well DMRs truly separate Early OC from Non-tumor without leakage. To test the hypothesis, the experimental design directly compares a standard "leaky" global selection model to the leakage-safe pipeline. By discovering DMRs in cfDNA, strictly inside each training fold, training a nested-CV elastic-net, and benchmarking against a label-permutation null, honest AUROC/AUPRC and clinically relevant sensitivity/specificity can be reported. To prevent overly optimistic bias from feature selection leakage, fold-internal discovery was used with nested cross-validation, a best practice recommended by Cawley & Talbot (5).

METHODS AND MATERIALS

Synthetic Dataset Generation and Simulation Design

A synthetic dataset was developed that mirrors the region-level cfDNA methylation counts found in plasma. This approach follows the structure and workflow described in Lu *et al* (2). In this method, plasma cfDNA is profiled by cfMeDIP-seq to produce region read counts across samples, and then normalized by library size (CPM or log-CPM). DMR-based feature selection and elastic-net classification using repeated train and test splits follows. Count noise was modeled using a Poisson or overdispersal process and ensured that the early-stage versus control class balance, which is important for detectability in plasma, was maintained. To prevent data leakage, all preprocessing and DMR selection are performed within the training folds, using label-permutation nulls.

Synthetic data was chosen over real cfDNA datasets as accurately benchmarking feature-selection leakage requires raw, high-dimensional data ($p>>n$) with known biological labels. Real raw cfMeDIP-seq patient datasets are typically restricted by strict privacy regulations, and publicly available supplementary data is often already pre-filtered for significant regions, making it impossible to properly test feature selection flaws.

Therefore, the simulation is an appropriate, methodologically faithful stand-in for cfMeDIP-seq cfDNA methylation data. The simulated data closely approximates real cfDNA methylome characteristics by replicating extreme sparsity, variable library sizes, and overdispersion ingrained in actual sequencing. The dataset can successfully capture the biological and technical variance seen in real patient liquid biopsies by modeling the baseline counts and sequencing noise through a gamma-Poisson framework. This makes it suitable for validating analysis pipelines without making clinical performance claims.

The dataset consisted of 90 samples: 30 Early OC and 60 Non-tumor, with 25 benign and 35 healthy individuals. Exactly 15,000 genomic regions (features) per sample were modeled. A signal of exactly 1,200 (8.0%) regions was set to differ between Early OC and Non-tumor samples, serving as the DMRs. The differential signal (effect size) was computed by multiplying the baseline counts of the OC samples in these regions by a multiplier uniformly distributed between 1.7 and 2.3. Raw region counts were generated using a gamma-Poisson framework (gamma shape 2.0, scale 1.0) scaled by log-normally distributed library sizes (mean = 16, standard deviation = 0.3), and then normalized to log counts per million (log-CPM).

Modeling Workflow and Hyperparameter Selection

To evaluate the predictive power of the cfDNA methylation signals without data leakage, a nested cross-validation (nested-CV) approach was used along with an

elastic-net logistic regression model.

To prevent data leakage, a feature selection was performed only on the training data within each fold. DMRs were identified using a Negative Binomial Generalized Linear Model (GLM) that incorporated library-size offsets. P-values were corrected for the false discovery rate using the Benjamini-Hochberg procedure. The top 300 DMRs based on q-value and log2 fold-change were then selected.

The log-CPM features that corresponded to the selected DMRs using a standard scaler were then standardized. The model was evaluated using a 5-fold repeated and stratified K-Fold cross-validation strategy (5 splits and 5 repeats). Inside each outer training fold, an internal 5-fold Stratified K-Fold loop was used to optimize the elastic-net hyperparameters via a grid search. The model utilized the saga solver, which has a maximum of 5,000 iterations; this optimizes for the Area Under the Receiver Operating Characteristic Curve (AUROC). The hyperparameter grid explored both the L1 ratio (0.1, 0.5, 0.9) and the inverse regularization strength, C (0.1, 1.0, 10.0). The best-performing model from this inner loop was then evaluated on the unseen data in the outer test fold. Out-of-fold predictions were generated by aggregating the predicted probabilities from these unseen test folds across all 25 cross-validation iterations using 5 splits x 5 repeats. The final AUROC and Area Under the Precision-Recall Curve (AUPRC) were then calculated using these pooled out-of-fold predictions.

IRB statement

Because this study relied exclusively on a simulated synthetic dataset, Institutional Review Board (IRB) approval and informed consent were not required.

RESULTS

Using fold-internal DMR discovery with nested-CV elastic-net on the cfMeDIP-like simulated cfDNA dataset, an AUROC = 0.938 (Figure 1) and AUPRC = 0.914 (Figure 2) (out-of-fold) were obtained. A leakage ablation showed that global (leaky) DMR selection inflated performance (near-perfect AUROC under label permutation), whereas the proper fold-internal pipeline returned a permuted AUROC ≈ 0.5, confirming leakage-safe evaluation.

To further evaluate the model's viability, diagnostic performance was assessed across the aggregated out-of-fold predictions from the 25 nested cross-validation iterations. At the optimal classification threshold (0.577), the leakage-safe pipeline achieved a sensitivity (recall) of 83.3% and a specificity of 91.7%. This performance, alongside an out-of-fold AUROC of 0.938 and an AUPRC of 0.914, demonstrates the model's ability to robustly identify tumor-specific methylation patterns despite the extreme sparsity of the data and a 1:2 class imbalance (Figure 3).

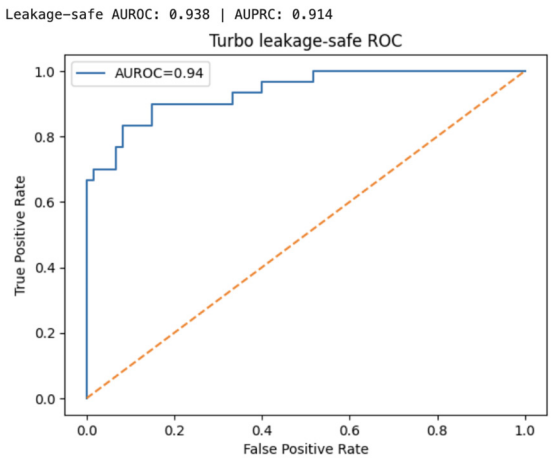

Figure 1. *Receiver Operating Characteristic (ROC) curve evaluating the leakage-safe pipeline on the simulated cfDNA dataset: (90 total samples: 30 Early OC, 60 Non-tumor). The x-axis represents the False Positive Rate (1 - specificity), and the y-axis represents the True Positive Rate (sensitivity). The curve is generated from the aggregated out-of-fold predicted probabilities of the nested-CV elastic-net model, achieving an AUROC of 0.938.*

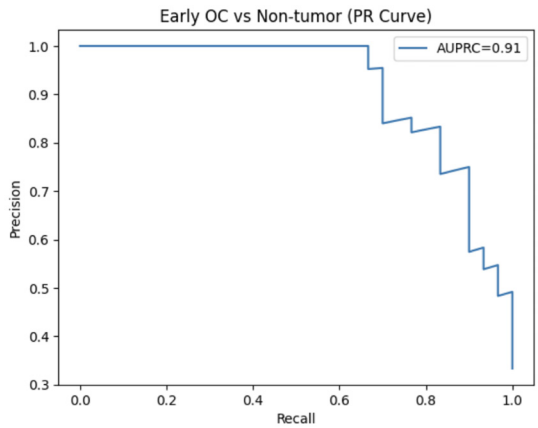

Figure 2. *Precision-Recall (PR) curve evaluating the same out-of-fold predictions on the simulated cfDNA dataset. The x-axis denotes Recall (sensitivity), and the y-axis denotes Precision (positive predictive value). Despite the 1:2 class imbalance in the data (30 Early OC to 60 Non-tumor), the model maintains high precision across most recall thresholds and yields an AUPRC of 0.914.*

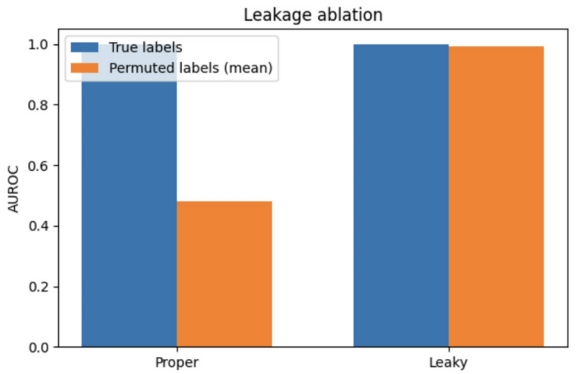

Figure 3. *Leakage ablation analysis comparing the proper fold-internal feature selection pipeline against a standard "leaky" global feature selection pipeline. The y-axis represents the Area Under the ROC Curve (AUROC). The blue bars show performance on true labels, while the orange bars show the mean performance over label permutations. The leaky pipeline artificially inflates the permuted AUROC to approximately 0.99 (an evaluation artifact), whereas the proper pipeline correctly restores the permuted AUROC to chance (approximately 0.50).*

DISCUSSION

These simulated findings demonstrate that under rigorous, leakage-safe conditions, early ovarian cancer methylation signals remain theoretically detectable despite the extreme sparsity characteristic of plasma cfDNA. The achieved sensitivity of 83.3% and specificity of 91.7% indicate that the nested-CV elastic-net pipeline successfully isolates true differentially methylated regions (DMRs) from high-dimensional background noise. Unlike global feature selection models that artificially inflate performance metrics through data leakage, this fold-internal approach provides a realistic assessment of biomarker viability in sparse, clinical-like datasets.

These results are consistent with previous studies that highlight the importance of rigorous machine learning validation in genomics. Feature selection leakage often causes inflated performance estimates, as noted by Cawley and Talbot (5). By simulating the cfMeDIP-seq workflows described by Lu *et al*. (2), this analysis shows that while cfDNA contains a detectable biological signal, the choice of computational method is just as important as the assay itself for avoiding overfitting.

The development of AI-driven liquid biopsy has advantages beyond computational rigor: it carries profound implications for health equity. Ovarian cancer's high mortality rate is largely driven by late-stage diagnoses resulting from the asymptomatic nature of the disease and the invasiveness of traditional diagnostics. Recent progress in molecular biology and computational genomics has enabled scalable, non-invasive approaches for early cancer detection. Robust computational pipelines support the implementation of liquid biopsy tools in a range of clinical environments, helping to close important gaps in early screening for women's health.

CONCLUSION

As the field moves toward AI-driven liquid biopsies, rigorous computational validation is just as critical as the biological assays themselves. By demonstrating the severe inflation caused by global feature selection and establishing a leakage-safe nested-CV pipeline, this study highlights the necessity of fold-internal DMR discovery. Because of this shift toward rigorous, internally validated machine learning, the field is moving closer to a clinically viable liquid biopsy that could detect ovarian cancer sooner, at a stage where therapy is most effective.

FUNDING SOURCES

The authors received no financial support for this research, authorship, and/or publication of this article.

CONFLICT OF INTEREST

The authors declare no conflicts of interest related to this work.

REFERENCES

1. National Cancer Institute. Cancer of the ovary - cancer stat facts. SEER. Available from: https://seer.cancer.gov/statfacts/html/ovary.html (accessed on 2026-01-22).
2. Lu H, Liu Y, Wang J, *et al*. Detection of ovarian cancer using plasma cell-free DNA methylomes. *Clin Epigenetics*. 2022; 14 (1): 64. https://doi.org/10.1186/s13148-022-01285-9
3. Galardi F, De Luca F, Romagnoli D, *et al*. Cell-free DNA-methylation-based methods and applications in oncology. *Biomolecules*. 2020; 10 (12): 1677. https://doi.org/10.3390/biom10121677
4. Moser T, Kühberger S, Lazzeri I, Vlachos G, Heitzer E. Bridging biological cfDNA features and machine learning approaches. *Trends Genet*. 2023; 39 (4): 315-326. https://doi.org/10.1016/j.tig.2023.01.004
5. Cawley G, Talbot N. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res*. 2010; 11: 2079-2107.