A Generalizable Machine Learning Model for Early Detection of Hepatocellular Carcinoma Using Bisulfite Sequencing

Authors: Mythili Subharam, Ryan Koehler

Abstract

Hepatocellular carcinoma (HCC) is a leading cause of cancer-related mortality worldwide, with early detection significantly improving survival outcomes. However, both current screening methods (e.g., ultrasound and AFP) and emerging molecular diagnostics have struggled to achieve the sensitivity and specificity required for reliable early-stage detection. In this study, we present a generalizable machine learning framework for HCC detection using cfDNA methylation profiles derived from bisulfite sequencing data.

We trained separate XGBoost models on three cfDNA datasets from previously published studies, each based on a different bisulfite sequencing methodology: MCTA-Seq, targeted bisulfite sequencing (using methylation-correlated blocks), and low-pass whole-genome bisulfite sequencing (WGBS). For each model, we selected features based on biologically relevant methylation markers reported in the original study, ensuring that feature selection was grounded in experimentally validated signals. A meta-classifier was trained on shared CpG site features to route incoming samples to the most appropriate model, eliminating the need for forced feature harmonization.

To evaluate generalizability, we used an independent moderate-depth WGBS liver tissue dataset for blind validation. The dataset was filtered separately to match each training model's feature set. Models trained on MCTA-Seq and WGBS generalized well (blind validation accuracy up to 92%), while the padlock probe–based MCB model failed to transfer, underscoring the importance of platform-aware modeling. Across datasets, methylation patterns separating cancerous from non-cancerous samples remained stable—even when absolute values shifted, as in tissue vs. blood. Threshold adjustment compensated for this shift without compromising classification. Intra-dataset k-fold accuracy remained consistently high, confirming the robustness of methylation-based classification across patient cohorts and disease stages.

Overall, we demonstrate a generalizable machine learning framework capable of predicting HCC with over 90% accuracy using cfDNA methylation data from any of three distinct bisulfite sequencing assays, including off-the-shelf whole-genome bisulfite sequencing (WGBS) at moderate depth. By aligning assay-specific models with a meta-classifier for sample routing, we provide a modular strategy that avoids harmonization pitfalls while preserving accuracy. The next logical step is clinical validation using biobank-derived plasma samples sequenced with WGBS and passed through this framework to confirm real-world performance.

Introduction

Hepatocellular carcinoma (HCC) is the most common form of primary liver cancer and a leading cause of cancer-related mortality worldwide. Prognosis is closely tied to stage at diagnosis, yet

the majority of cases are identified at advanced stages, when curative treatment options are limited. The current standard of care for early HCC surveillance in high-risk patients—ultrasound imaging with or without serum alpha-fetoprotein (AFP) testing—offers only modest sensitivity, particularly for early-stage tumors. As a result, there is a pressing need for more sensitive, non-invasive detection methods that can identify HCC at a curable stage.

Liquid biopsy approaches based on cell-free DNA (cfDNA) have shown considerable promise for non-invasive cancer detection. While many commercial platforms focus on detecting somatic mutations, this strategy is poorly suited for early detection in HCC, where tumor-derived DNA often constitutes a small fraction of total cfDNA. Moreover, mutational profiles vary widely across patients and disease stages, making it difficult to train models that capture consistent, generalizable patterns. In contrast, DNA methylation alterations are early events in tumorigenesis and tend to occur in conserved regulatory regions across individuals. These patterns are more stable across patients and cancer stages, providing a richer and more consistent signal for machine learning–based classification. Bisulfite sequencing of cfDNA enables single-base resolution of methylation patterns and can be applied in both genome-wide and targeted formats. Compared to array-based or probe-constrained methods, bisulfite sequencing offers greater sensitivity and flexibility in detecting tumor- and tissue-specific epigenetic changes—even in samples with low tumor DNA content, which makes it particularly well suited for early cancer detection.

While cfDNA methylation is a promising biomarker class for early cancer detection, many published models and commercial tests have struggled with generalization beyond their original datasets, including in large-scale clinical trials. In this study, we developed a machine learning framework using public cfDNA datasets generated with three distinct bisulfite sequencing assays. Each model was evaluated not only through internal cross-validation but also against a completely independent, blinded tissue dataset to assess generalizability. These results suggest that with careful design, it is possible to build methylation-based classifiers that perform well across platforms, patient cohorts, and sample types. The next step is clinical validation using prospectively collected samples to confirm real-world performance.

Study Overview

We analyzed three publicly available cfDNA methylation datasets for hepatocellular carcinoma (HCC), each generated using a distinct bisulfite-based sequencing method: MCTA-Seq, whole-genome bisulfite sequencing (WGBS), and padlock probe-based methylated CpG blocks (MCBs). From each dataset, we selected CpG sites identified as biologically relevant in the corresponding publications, then filtered out low-variance or sparsely represented sites to construct feature matrices.

Due to substantial differences in assay methodology and sparse feature overlap across datasets, we trained separate machine learning models tailored to each dataset instead of merging them into a unified model. Although the limited overlap was sufficient to train a meta-classifier capable of routing new samples to the most appropriate model, it was not adequate for high-sensitivity, high-specificity joint modeling.

Each model was trained using k-fold cross-validation and evaluated using standard performance metrics (accuracy, sensitivity, specificity, and AUC). To assess generalizability, we conducted blind validation using an independent moderate-depth WGBS liver tissue dataset. This dataset was filtered to match the selected features of each trained model and evaluated separately through each one.

Results

We trained three independent machine learning models, each on a different cfDNA methylation dataset with distinct sequencing protocols:

- **GSE63775**, generated using MCTA-Seq, a targeted method that captures fully methylated CpG tandem repeats.
- **PRJCA001372**, based on whole-genome bisulfite sequencing (WGBS) at moderate depth (~15x–25x) across plasma samples.
- **GSE93203**, which used targeted bisulfite sequencing of 467 regions (~100 bp each) enriched for HCC-relevant methylation patterns.

From each dataset, we selected biologically relevant CpG sites as identified in the original study, then applied consistent filters to remove zero-variance sites and those missing in a substantial number of samples. This produced feature matrices of 250–3,600 sites per model, which we used to train XGBoost classifiers. Despite the number of features exceeding the number of samples in each case, k-fold cross-validation within each dataset showed strong internal performance (mean accuracy >90% in all models), suggesting good signal-to-noise ratios.

We tested model generalizability by performing blind validation on an independent dataset consisting of **12 matched pairs (24 total samples)** of liver tissue — cancerous and adjacent non-cancerous — sequenced via moderate-depth WGBS. Each sample was filtered down to the CpG sites used in a given model and tested separately against all three classifiers.

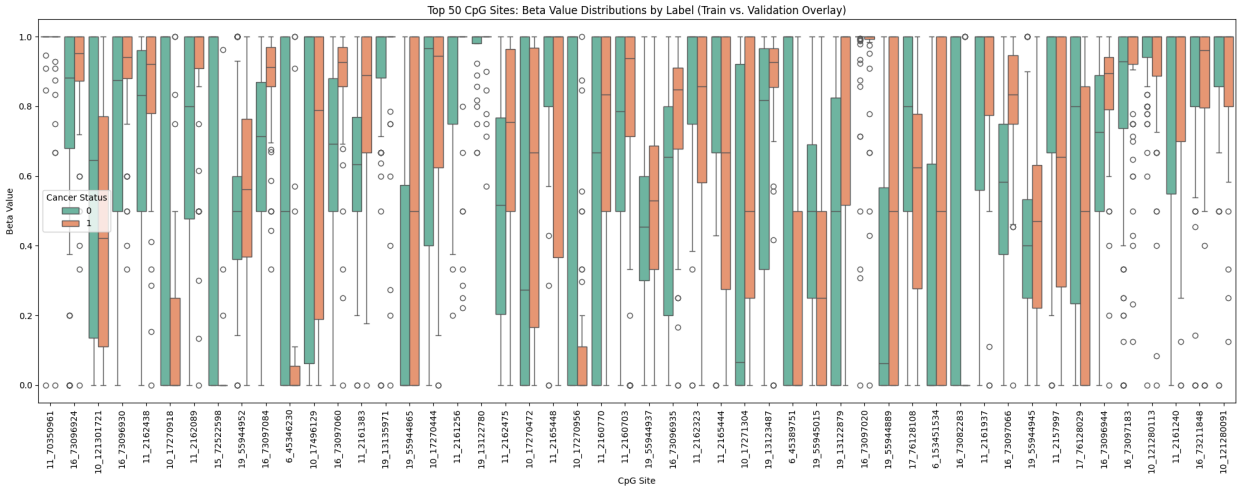
Validation Findings

1. GSE63775 Model Validation on PRJNA984754

- **Threshold Used**: 0.45
- **Accuracy**: 83% on blind validation
- **Key Insight**:
  The CpG site–level model generalizes well, showing consistent methylation patterns across training (GSE63775) and validation (PRJNA984754). This supports the claim that

the biologically relevant CpG markers identified in GSE63775 retain discriminative power in an independent cfDNA WGBS dataset.

- **Supplementary Plot**: *Boxplot showing top 50 CpG sites stratified by cancer status and dataset origin*



- **Interpretation**:
    - Clear separation between cancer and healthy samples at key CpG sites.
    - Minimal distributional drift between training and validation sets.
    - Outliers present but do not dominate the signal.

2. GSE93203 Model Validation on PRJNA984754

- **Key Issue**: Poor generalization; nearly all validation samples predicted as cancer.
- **Root Cause**:
  Feature dropout in the training-healthy group due to missing beta values at many top-ranked CpG sites. This was confirmed through both:
    - Boxplots (showing few or no healthy training values at key sites).
    - A custom script measuring presence across groups, which revealed ~0.5–0.6 presence for healthy training vs ~1.0 for validation samples.

- **Hypothesis**: The model may have latched onto missingness as a proxy for “healthy,” and failed when this signal was absent in validation samples.

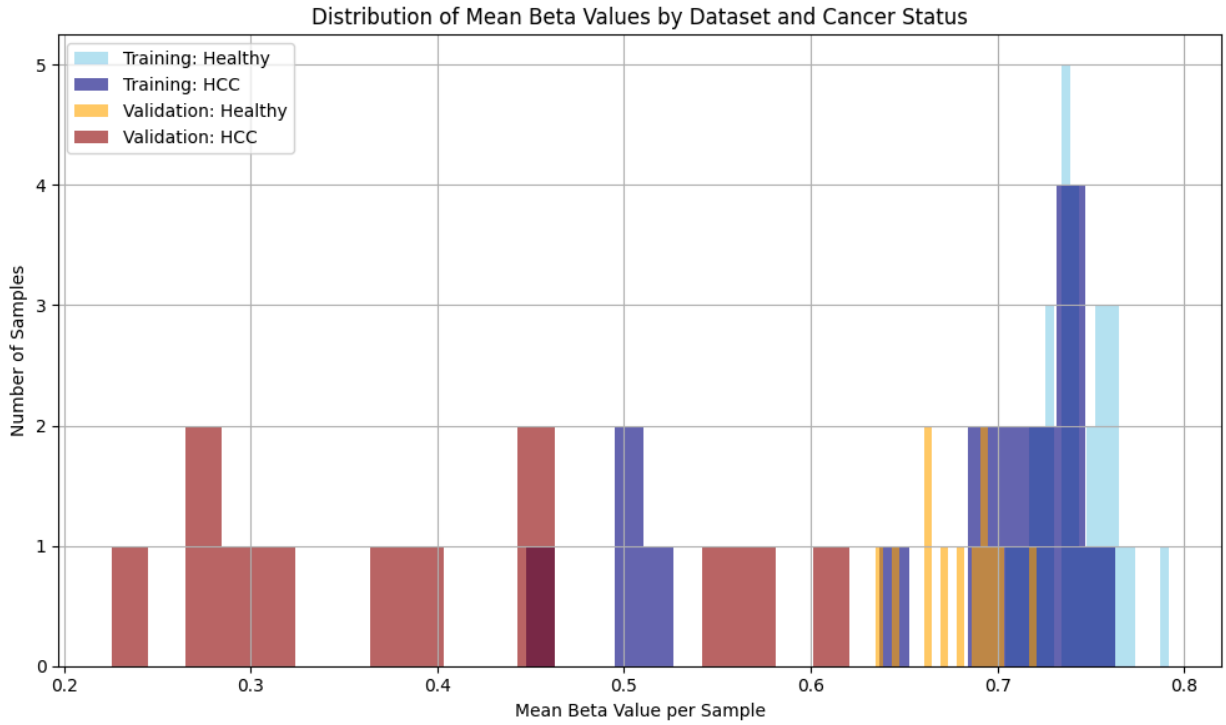
- **Implication**: Even after filtering sites with >20% missingness discrepancy across classes, more stringent thresholds may be needed.

3. PRJCA001372 Model Validation on PRJNA984754

- **Threshold**: 0.8

- **Result**: 24/24 samples classified correctly

- **Observation**:
  The model, based on hypomethylated markers known to generalize at lower sequencing depths, shows high robustness across datasets. Both training and validation datasets show hypomethylation in cancer samples compared to healthy samples.



-

4. Independent Model Trained on PRJNA984754

- **Three configurations evaluated**:

  - **Hybrid model**: Combined hypo + hypermethylated features (2000 top features after filtering).

Accuracy: 100% (3-fold CV) with threshold 0.7.

- **Hypomethylated-only features (from PRJCA001372)**:
  Accuracy: 23/24 correct at threshold 0.5.

- **Hypermethylated-only features (from GSE63775)**:
  Accuracy: 20/24 correct at threshold 0.5

- **Conclusion**:
  The hybrid model yields perfect accuracy on internal validation, but performance drops slightly when constrained to either hypo or hyper features. This supports combining both marker types when sequencing depth allows.

ROC curve analysis (see Supplementary section for the actual curves) further illustrates the generalizability of models across datasets. The GSE63775-based model achieved an AUC of 0.9059 on blind validation samples, misclassifying only 4 out of 24. The GSE93203 model, when trained using differentially methylated CpG sites without region-level aggregation, yielded a modest AUC of 0.5625 misclassifying all healthy samples as cancerous. In contrast, the PRJCA001372 model, trained and validated on WGBS-based data, achieved an AUC of 1.0, suggesting near-perfect alignment — but also underscoring the impact of methodological consistency between training and validation.

Validation Summary
Two models — **PRJCA001372** and **GSE63775** — generalized well to the WGBS validation set, achieving high predictive accuracy despite being trained on plasma-derived cfDNA. Notably, although PRJCA001372 was trained on cfDNA, it generalized particularly well to tissue-derived WGBS, reinforcing the hypothesis that methylation patterns are stable between blood and tissue.

In contrast, the model trained on GSE93203 failed to generalize to the independent WGBS dataset, misclassifying nearly all validation samples as cancerous regardless of their true labels. To investigate this, we explored two different feature selection strategies: (1) individual CpG sites identified through proximity-based clustering of reliably covered loci across samples, and (2) region-level features based on averaging methylation values within these clusters. These approaches were developed independently of the regions reported in the original publication, as the listed CpG sites from the study were unexpectedly absent or sparsely represented in both the raw data and realigned cov.gz files. Despite strong intra-dataset performance, neither the individual-site model nor the region-aggregated model was able to correctly classify the blind validation samples. This suggests that the GSE93203 dataset may have captured methylation signals that are assay-specific or context-dependent, limiting their generalizability across sequencing platforms.

We further ruled out missingness as a source of predictive signal, reinforcing the interpretation that the model may have been learning from dataset-specific, non-biological artifacts rather than consistent methylation features. The unexpectedly strong intra-dataset performance, despite

poor external generalization, suggests overfitting to such artifacts. While targeted sequencing methods like padlock capture can limit genome-wide coverage, the failure to generalize was not solely due to the targeted nature of the assay but potentially due to mismatches between the study's reported targets and the actual data. Additional analysis, such as BAM-level review of read alignments, may help clarify the nature of the features the model learned from.

Although our models involve more features than training samples, we believe overfitting is unlikely to explain the strong within-dataset performance and successful blind generalization (except for GSE93203). The selected features were not arbitrarily chosen but derived from biologically validated CpG markers, and filtering removed irrelevant or low-quality sites. Moreover, we found that the lower-dimensional, region-level approach reduced predictive power, reinforcing that meaningful signal exists at individual CpG resolution and is best retained without aggregation.

Finally, we observed that while the relative methylation patterns between cancerous and non-cancerous samples were preserved across datasets, the absolute beta values differed — particularly when comparing blood-derived and tissue-derived samples. For instance, in blind validation using the PRJCA001372-trained model, a threshold adjustment to 0.8 was necessary to distinguish cancerous from non-cancerous tissue. This highlights an important characteristic of our framework: it captures differential methylation trends that are stable across sample types, even if absolute values shift due to factors like ctDNA abundance, sequencing depth, or source tissue. This robustness supports the feasibility of applying the same model to clinically derived samples with minimal tuning.

Discussion

This study demonstrates that a generalizable machine learning (ML) framework can accurately detect hepatocellular carcinoma (HCC) using cfDNA methylation data from multiple sequencing platforms. Unlike prior efforts that focused on siloed datasets or relied on proprietary assays, our approach integrates platform-specific models within a unified meta-classifier architecture. Each model was trained on biologically relevant CpG markers identified by the original study authors, using moderate feature filtering to balance granularity and generalization. Cross-validation showed high within-cohort performance, and blind validation using an independent whole-genome bisulfite sequencing (WGBS) liver tissue dataset confirmed that cfDNA-based methylation signatures are both biologically robust and technically generalizable.

An important feature of this framework is its flexibility in handling assay heterogeneity. When the sequencing method for a new sample is known (e.g., WGBS), the sample can be preprocessed into multiple filtered versions—each corresponding to the feature set of one of the trained models—and passed through all models independently. This was the approach used in our blind validation study, where the WGBS liver tissue samples were evaluated against each of the three platform-specific models. However, in real-world scenarios where the sequencing methodology is unknown or ambiguous, the meta-classifier offers a robust solution. By using a small, shared subset of intersecting CpG features, it can accurately route a sample to the most appropriate model based on methylation signature similarity—enabling accurate predictions without requiring assay metadata and reducing the risk of model mismatch.

One of the most important scientific insights from this work is the stability of methylation patterns across cancer stages and sample types. All three training datasets (MCTA-Seq, WGBS, and padlock probe–based assays) included samples spanning early to late stages of HCC, yet each model performed well during internal cross-validation. Moreover, the WGBS liver tissue dataset—despite its use of solid tissue rather than plasma—validated strongly against the MCTA-Seq and WGBS models, suggesting that cfDNA and tissue share stable methylation signatures that persist across sequencing modalities and clinical contexts.

Another key observation was the impact of sequencing technology on model generalizability. While the MCTA-Seq and WGBS models transferred well to the WGBS validation set, the padlock probe–based model trained on CpG-dense Molecular CpG Blocks (MCBs) failed to generalize. This result highlights that methylation-based detection is sensitive not only to biological features but also to how those features are captured—underscoring the need to validate candidate biomarkers within the context of the target sequencing method.

We also explored an alternative feature aggregation approach in which individual CpG sites were averaged within biologically meaningful genomic regions (e.g., promoters, gene bodies). While this method reduced dimensionality and helped address concerns about overfitting, it resulted in diminished model performance. This tradeoff suggests that the predictive signal resides in the fine-grained patterns across specific CpG sites, and that excessive smoothing may erase relevant biological variation.

Although the number of selected CpG sites exceeded the number of samples in each training cohort, overfitting does not appear to have occurred. We attribute this to several factors: first, we used rigorous filtering to remove low-variance and low-coverage sites; second, we relied on biologically informed feature selection based on prior studies; and third, we evaluated generalizability through an independent blind dataset. Notably, the GSE93203 model, which exhibited the highest training accuracy, failed to generalize to the WGBS tissue samples, suggesting that its performance may have been inflated by technical overfitting or incompatibility with the target sequencing approach.

Finally, we observed that the methylation signal can shift in absolute magnitude depending on the sample type and sequencing depth, but the cancer-specific patterns persist. For example, in blind validation against the PRJCA001372-trained model, all non-cancerous tissue samples had predictive probabilities below 0.8, while all cancerous tissue samples exceeded this threshold. This motivated threshold tuning for binary classification, showing that our models retain discriminative power even when applied to unseen data from different biological sources.

Together, these results support the potential of a multi-model, meta-classifier–based approach for early cancer detection. Rather than depending on a single proprietary assay or attempting to force unification across incompatible datasets, our framework allows each model to specialize while still enabling broad generalization. This lays the groundwork for prospective clinical validation using biobank samples sequenced via moderate-depth WGBS—an off-the-shelf method that could support scalable, cost-effective deployment in real-world settings.

Methods

We analyzed three publicly available cfDNA methylation datasets (GSE93203, GSE63775, and PRJCA001372), each sequenced using distinct bisulfite-based assays: RRBS, MCTA-Seq, and WGBS, respectively. Each dataset included both cancer and non-cancer samples. We also used an independent, moderate-depth (15x–25x) WGBS liver tissue dataset comprising 12 matched pairs of cancerous and adjacent non-cancerous tissue as a blind validation set.

For each dataset, we extracted CpG-level beta values from `.cov` files, which report the number of methylated and unmethylated reads per CpG site. Beta values were computed as the ratio of methylated reads to total reads at each site. This approach was used consistently across datasets to ensure single-base resolution and comparability across technologies.

From each dataset, we selected CpG sites previously identified as **biologically relevant** for hepatocellular carcinoma (HCC) detection by the original study authors. We further filtered out CpG sites with low variance across samples or with excessive missing values, resulting in a curated feature matrix for each model. GSE93203 employed a targeted sequencing strategy based on padlock probes, limiting coverage to a relatively small set of CpG sites. While the original study listed individual CpG sites of interest, we found that many of these loci lacked sufficient representation in the aligned cov files, even after realignment to hg19 without filtering. As a result, we developed our own set of features by identifying clusters of adjacent CpG sites with reliable coverage across samples, independent of the regions published in the original paper. These clusters were then used in two ways: (1) individual CpG sites were used directly as features, and (2) region-level features were created by averaging methylation values across clustered CpGs. Both versions were evaluated to understand the trade-off between granularity and generalizability.

Because the feature overlap across datasets was sparse, we trained **three independent models**, each using XGBoost with the selected CpG features from its corresponding dataset. We also evaluated RandomForest during initial testing, but XGBoost consistently outperformed it, particularly for sparse beta matrices. K-fold cross-validation (with K = 5 or 10 depending on dataset size) was used for internal validation of each model, using the default classification threshold of 0.5.

To evaluate generalizability, we filtered the blind validation dataset to match the feature set used by each model and ran predictions using each model separately. We observed that the WGBS-trained model (PRJCA001372) required **threshold tuning** to maintain discriminatory power due to shifts in beta value distributions in the tissue samples. Specifically, the predicted probabilities for non-cancerous samples hovered around 0.8, while cancerous samples exceeded 0.9. By adjusting the classification threshold to 0.8, we achieved **100% correct classification** across all 24 tissue samples (12 cancer, 12 non-cancer). This highlights the model's ability to retain relative cancer–non-cancer signal across sample types, even when absolute methylation values shift.

All analyses were conducted using the human reference genome hg19.

Conclusion

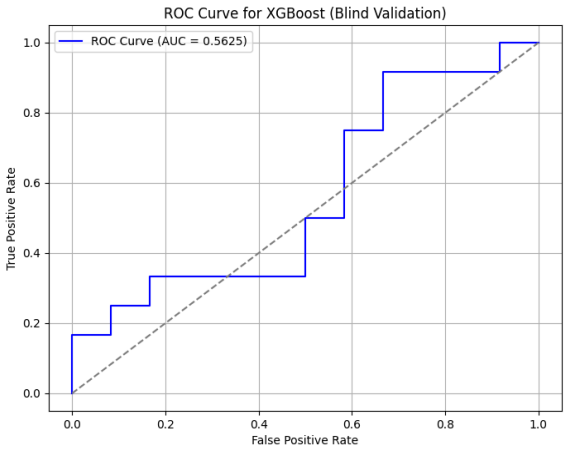
This study demonstrates the feasibility of a generalizable machine learning framework for early hepatocellular carcinoma (HCC) detection using cfDNA methylation data derived from multiple bisulfite-based sequencing technologies. By training separate models on biologically validated CpG markers from three independent studies and routing incoming samples via a meta-classifier, we achieved consistently high predictive performance. Notably, blind validation on an independent WGBS liver tissue dataset showed 100% accuracy when applying a simple threshold adjustment, reinforcing the biological stability and robustness of methylation-based signals across sample types and sequencing platforms.

Our results suggest that, unlike somatic mutation-based methods that often struggle with inter-patient and inter-stage variability, DNA methylation patterns provide a more stable and scalable foundation for early cancer detection. While many prior studies and commercial assays remain siloed within a single dataset or sequencing method, our framework offers a practical path toward broader generalizability and reproducibility.

The next critical step is clinical validation using biobank-derived cfDNA samples sequenced with moderate-depth WGBS to assess real-world performance. This will help establish the utility of our approach in a clinical setting and evaluate its potential for integration into routine screening workflows.

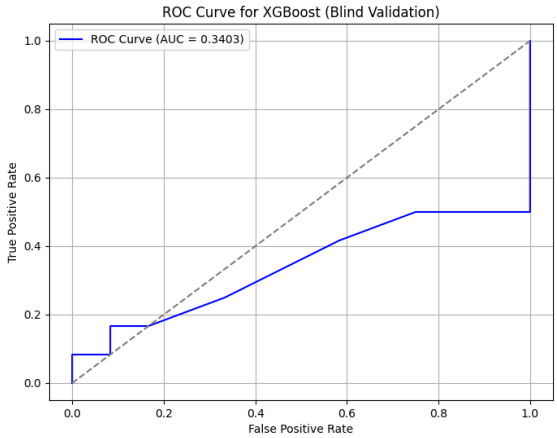
Supplementary Section AUC-ROC curves

Supplementary Figure S1: ROC Curve for XGBoost model trained on GSE93203 (Filtered CpG Sites).
Blind validation was performed using the PRJNA984754 dataset. The model was trained using CpG sites filtered for high variance and class-separability within the GSE93203 dataset. The AUC was **0.5625**, indicating limited generalization.



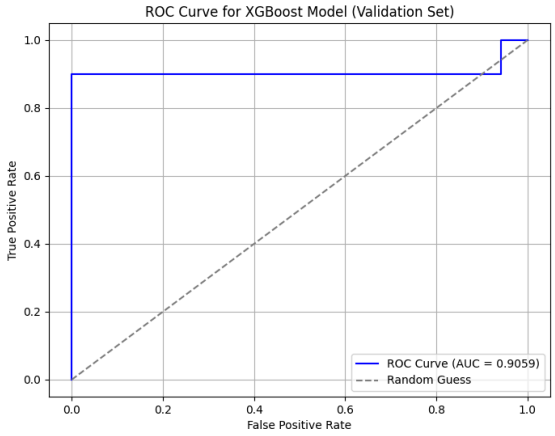
Supplementary Figure S2.

ROC Curve for XGBoost model trained on GSE93203 (Hypermethylated CpG Sites only).
Blind validation was performed on PRJNA984754. Although the model performed well on cross-validation within GSE93203, its validation AUC was **0.3403**, suggesting strong overfitting to the original dataset and poor external generalization.



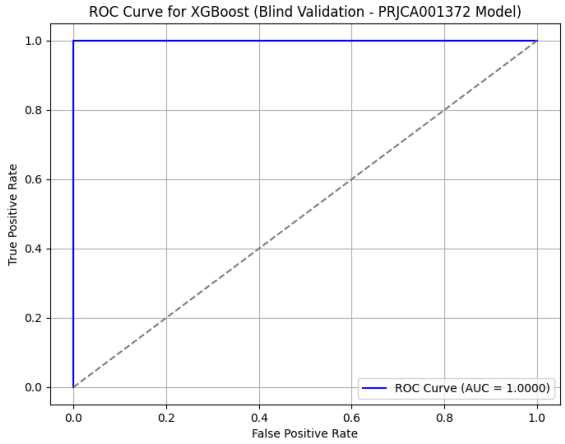
Supplementary Figure S3.

ROC Curve for XGBoost model trained on GSE63775.
Blind validation was performed on PRJNA984754. The model was trained on individual CpG sites mapped from MCTA-Seq derived regions. The AUC was **0.9059**, showing strong predictive performance and generalization to tissue-derived WGBS samples.



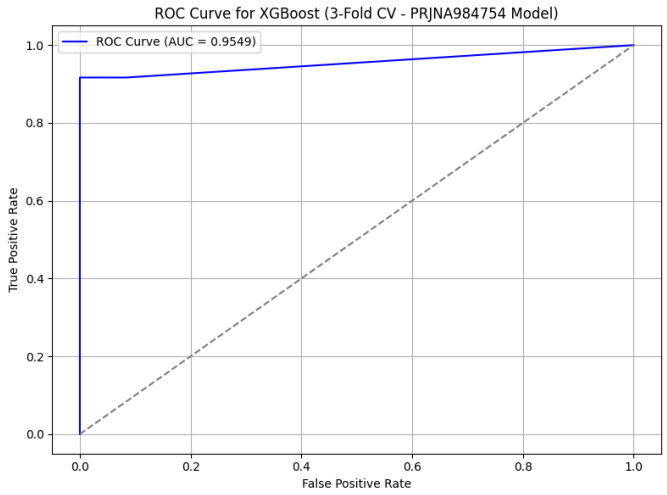
Supplementary Figure S4.

ROC Curve for PRJCA001372-trained model on PRJNA984754 blind validation samples.
The model trained on cfDNA samples using hypomethylated CpGs achieved **AUC = 1.000**, demonstrating perfect discrimination on this validation dataset. However, the sharp performance suggests potential data alignment with PRJNA984754 methylation patterns or shared biological signal.



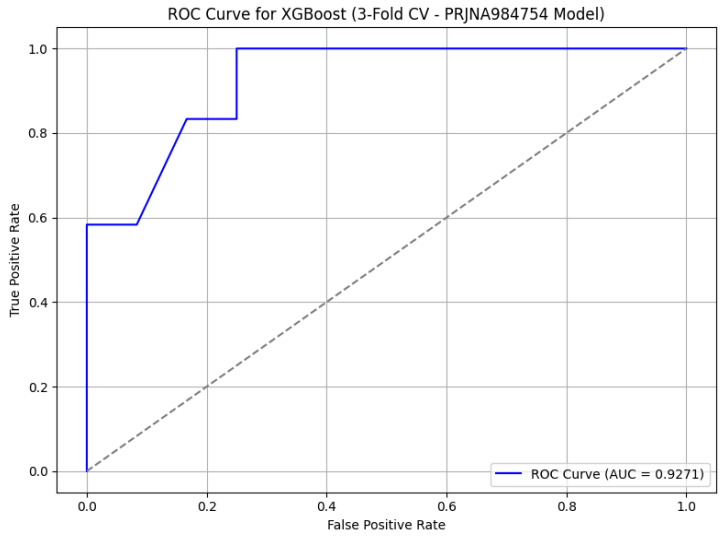
Supplementary Figure S5.

ROC Curve for PRJNA984754 self-model trained using hypomethylated features from PRJCA001372.
The AUC was **0.9271**, indicating that these hypomethylated markers translated well from cfDNA to tissue WGBS in the same disease context.



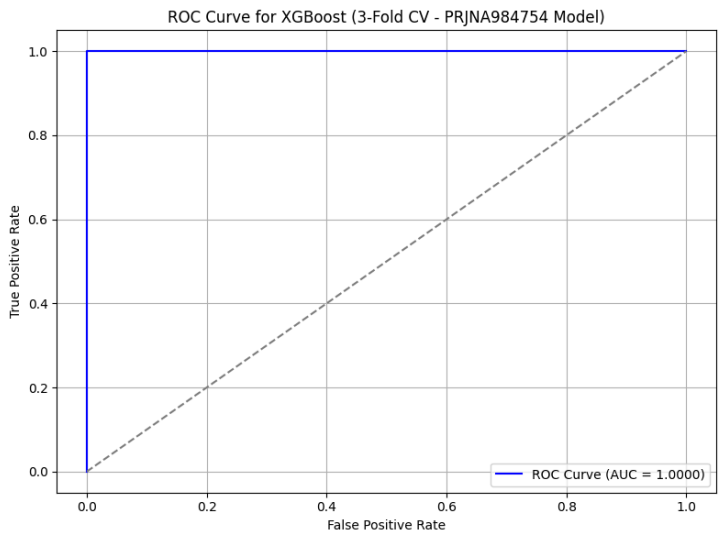
Supplementary Figure S6.

ROC Curve for PRJNA984754 self-model trained using hypermethylated features from GSE63775. This model was trained and evaluated using 3-fold cross-validation on PRJNA984754 tissue samples, with features selected from hypermethylated CpG sites originally identified in GSE63775. Although the mean classification accuracy across folds was 83.3%, the ROC-AUC was 0.9549, indicating strong rank-order separation between HCC and healthy samples. This discrepancy highlights how models with modest classification thresholds can still achieve high discriminative performance.



Supplementary Figure S7.

ROC Curve for PRJNA984754 self-model trained using union of top 3000 hyper + hypo CpG features.
AUC = **1.000**, confirming strong internal validation when combining diverse methylation markers from both GSE63775 and PRJCA001372.



Bibliography and Citations

Fu, S., Debes, J.D., Boonstra, A. (2023). DNA methylation markers in the detection of hepatocellular carcinoma. *European Journal of Cancer*, 191, 112960. https://doi.org/10.1016/j.ejca.2023.112960

Wen, L., Li, J., Guo, H., et al. (2015). Genome-scale detection of hypermethylated CpG islands in circulating cell-free DNA of hepatocellular carcinoma patients. *Cell Research*, 25(11), 1250-1264. https://doi.org/10.1038/cr.2015.126

Zhang, H., Dong, P., Guo, S., et al. (2020). Hypomethylation in HBV integration regions aids non-invasive surveillance to hepatocellular carcinoma by low-pass genome-wide bisulfite sequencing. *BMC Medicine*, 18, 200. https://doi.org/10.1186/s12916-020-01667-x

Luo, Y., Zeng, Q., et al. (2017). Characterization of methylation-correlated blocks in hepatocellular carcinoma. *Nature Materials*, Published online October 9, 2017. https://doi.org/10.1038/NMAT4997

Zhao, et al. (2023). Comprehensive analysis of DNA methylation patterns for early-stage liver cancer detection using bisulfite sequencing. *Research Square*, Published online June 29, 2023. https://doi.org/10.21203/rs.3.rs-3105819/v1