BMC Cancer

https://doi.org/10.1186/s12885-026-15720-0

Article in Press

Multimodal analysis of cell-free DNA to improve early detection of gastric cancer

Received: 10 March 2025

Accepted: 6 February 2026

Published online: 23 February 2026

Cite this article as: Vo D.L., Huynh L.A.K., Vo D.H. *et al.* Multimodal analysis of cell-free DNA to improve early detection of gastric cancer. *BMC Cancer* (2026). https://doi.org/10.1186/s12885-026-15720-0

Duy Long Vo, Le Anh Khoa Huynh, Dac Ho Vo, Thi Hue Hanh Nguyen, Thi Tuong Vi Van, Giang Thi Huong Nguyen, Thuy Nguyen Doan, Viet Hai Nguyen, Quang Dat Tran, Quang Thong Dang, Vu Tuan Anh Nguyen, Le Minh Quoc Ho, Thi Phuong Dung Ha, Thi Ngoc Dung Dang, Pham Thanh Nhan Nguyen, Khac Tien Nguyen, Van Chien Ho, Thi Loc Le, Thi Hong Nhung Nguyen, Ngoc Hieu Tu, Thanh Son Tran, Thanh Xuan Jasmine, Thi Loan Vo, Thi Huong Thoang Nai, Thuy Trang Tran, My Hoang Truong, Ngan Chau Tran, Thanh Cong Nguyen, Thi Truc Nguyen, Bao Toan Le, Van Phong Tang, Thi Tu Nguyen, Anh Tuan Nguyen, Hoang Giang Vu, Thi Phan, Thi Ngoc Tien Nguyen, Hoang Anh Cao, Trong Hieu Nguyen, Lan N. Tu, Hoa Giang, Minh Duy Phan, Hoai-Nghia Nguyen, Van Thien Chi Nguyen & Le Son Tran

Multimodal analysis of cell-free DNA to Improve Early Detection of Gastric Cancer

Duy Long Vo[1,2#], Le Anh Khoa Huynh[3#], Dac Ho Vo[3], Thi Hue Hanh Nguyen[3], Thi Tuong Vi Van[3], Giang Thi Huong Nguyen[3], Thuy Nguyen Doan[1], Viet Hai Nguyen[1], Quang Dat Tran[1], Quang Thong Dang[1], Vu Tuan Anh Nguyen[2], Le Minh Quoc Ho[1], Thi Phuong Dung Ha[4], Thi Ngoc Dung Dang[4], Pham Thanh Nhan Nguyen[5], Khac Tien Nguyen[6], Van Chien Ho[6], Thi Loc Le[7], Thi Hong Nhung Nguyen[7], Ngoc Hieu Tu[8], Thanh Son Tran[8], Thanh Xuan Jasmine[9], Thi Loan Vo[9], Thi Huong Thoang Nai[9], Thuy Trang Tran[9], My Hoang Truong[9], Ngan Chau Tran[9], Thanh Cong Nguyen[10], Thi Truc Nguyen[10], Bao Toan Le[11], Van Phong Tang[11], Thi Tu Nguyen[3], Anh Tuan Nguyen[3], Hoang Giang Vu[3], Thi Van Phan [3], Thi Ngoc Tien Nguyen [3], Hoang Anh Cao [3], Trong Hieu Nguyen[3], Lan N Tu[3], Hoa Giang[3], Minh Duy Phan[3], Hoai-Nghia Nguyen[3], Van Thien Chi Nguyen[3*], Le Son Tran[3*]

[1]Department of Gastrointestinal Surgery, University Medical Center, University of Medicine and Pharmacy at Ho Chi Minh City, Vietnam

[2]Department of General Surgery, Faculty of Medicine, University of Medicine and Pharmacy at Ho Chi Minh City, Vietnam

[3]Medical Genetics Institute, Ho Chi Minh, Vietnam

[4]Ha Noi Medical University Hospital, Ha Noi, Vietnam

[5]Da Nang Oncology Hospital, Da Nang, Vietnam

[6]Nghe An Oncology Hospital, Nghe An, Vietnam

[7]Thai Nguyen Central General Hospital, Thai Nguyen, Vietnam

[8]Buon Ma Thuot Medical University, Buon Ma Thuot, Vietnam

[9]Medic Medical Center, Ho Chi Minh, Vietnam

[10]Military Medical Hospital 175, Ho Chi Minh, Vietnam

[11]Can Tho Oncology Hospital, Can Tho, Vietnam

[#]Duy Long Vo and Le Anh Khoa Huynh contributed equally to this study.

*Correspondence: chinguyen@genesolutions.vn and leson1808@gmail.com

A short running head (40 characters)

ctDNA based Assay for Gastric Cancer Detection

Word count: 5253 words

ABSTRACT

Background:

Gastric cancer remains a global health challenge due to the difficulty of detecting it early in asymptomatic, high-risk populations. Current invasive diagnostic methods are impractical for widespread screening. Liquid biopsy using circulating tumor DNA (ctDNA) shows promise, but early detection is hindered by the low abundance and heterogeneity of ctDNA.

Methods:

We developed a multimodal cfDNA assay integrating methylation, fragmentomic, and hotspot mutation profiling from a single blood draw to detect gastric cancer-specific molecular signatures. Using these signatures, a machine-learning model was trained on a discovery cohort of 110 nonmetastatic GC patients and 119 healthy controls, then validated on an independent cohort of 58 patients and 65 controls.

Results:

The ensemble model achieved an AUC of 0.87 (95% CI: 0.80-0.93), with 70.7% sensitivity and 92.3% specificity for detecting nonmetastatic GC. Incorporating hotspot mutation profiling increased overall sensitivity to 75.9% without affecting specificity. Compared to a previous multi-cancer model, our ensemble model showed improved sensitivity across all stages, particularly for early-stage GC (72.7% vs. 36.4%).

Conclusions:

This multimodal cfDNA assay provides a minimally invasive and effective strategy for early GC detection, making it a potential screening tool for high-risk populations.

Mini abstract (30 words)

This study presents a novel multimodal cfDNA assay that combines methylation, fragmentomic, and hotspot mutation profiling, achieving 75.9% sensitivity and 92.3% specificity for early gastric cancer detection.

Keywords: gastric cancer, cfDNA, hotspot mutations, methylation and fragmentomic.

INTRODUCTION

Gastric cancer (GC) is the second leading cause of cancer-related morbidity and mortality worldwide, with a particularly high prevalence in Asian countries [1]. In Vietnam, GC ranks among the top five most common cancer types, representing 9.0% of all cancer cases according to Globocan 2022 [2]. Alarmingly, the majority of GC cases in Vietnam are diagnosed at advanced stages, contributing to high mortality rates and poor 5-year survival outcomes [3].

Early diagnosis and prompt treatment markedly enhance survival outcomes for patients with GC. The current GC detection method is upper endoscopy, though sensitive for early GC detection, requires skilled operators and complex procedures, and has limitations in terms of cost, invasiveness, and patient compliance. Other non-invasive methods include *Helicobacter pylori* serology and serum pepsinogen testing, serve only as biomarkers for GC risk and are ineffective in population-based screening programs [4]. Therefore, developing a cost-effective, non-invasive diagnostic method for early GC detection with high sensitivity is critically important. Such an advancement could significantly reduce mortality rates associated with GC.

Recent advances in genetic testing have led to the development of liquid biopsy, particularly the analysis of plasma cell-free nucleic acids, as a sensitive and cost-efficient tool for early cancer detection, staging, and treatment monitoring [5]. Liquid biopsy offers several advantages over traditional tissue biopsy, including reduced invasiveness and the ability to capture tumor heterogeneity. This technique allows for the detection of cancer-specific biomarkers, such as circulating tumor cell-free DNA (ctDNA), from body fluids, providing a promising avenue for cancer diagnosis and screening. Studies have shown that ctDNA-based liquid biopsies can effectively detect early cancer, as ctDNA contains mutations, epigenetic methylation changes, and unique fragmentomic profiles linked to cancer onset and progression [5, 6]. For instance,

The CancerSEEK test achieved 72% sensitivity and 99% specificity for detecting GC based on profiling 61 hotspot mutations in plasma cfDNA [7]. Similarly, Ren et al. (2022) reported a cfDNA methylation panel with 153 biomarkers, achieving 44% sensitivity for stage I and 59% for stage II GC at a specificity of 92% [8]. Liu et al. (2020) reported the CCGA study in using methylation profile to detect more than 50 types of cancers, including GC. The study demonstrated 99.3% specificity and 67.3% sensitivity for stage I–III cancers (CI: 60.7% to 73.3%) across a pre-specified panel of 12 cancer types [9]. Together, these studies suggest that distinct genetic and epigenetic signatures may serve as potential biomarkers for detecting ctDNA in the blood of GC patients.

Despite these advances, the clinical application of cfDNA-based tests for GC remains challenging due to the low abundance of heterogeneity derived ctDNA in the bloodstream [10]. Previously, we developed the SPOTMAS assay, which utilizes the methylation and fragmentomic profiles of ctDNA for multi-cancer detection from a single blood sample including GC [6]. However, this assay primarily targeted biomarkers common to multiple cancer types, without specifically focusing on GC-related markers. In this proof-of-concept study, our objective is to comprehensively characterize and identify the unique methylation and fragmentomic signatures of cfDNA associated with GC to develop an ensemble machine learning classification model. Additionally, we evaluate the potential of integrating GC associated hotspot mutations to enhance detection sensitivity.

MATERIALS AND METHODS

Patient enrollment

This study recruited 168 gastric cancer (GC) patients diagnosed with non-metastatic tumors (stages 0-IIIA) and 184 healthy subjects, divided into a Discovery cohort and a Validation cohort.

The Discovery cohort included 110 GC patients and 119 healthy controls, while the Validation cohort comprised 58 GC patients and 65 healthy controls (**Figure 1, Figure S1** and **Table S1**). To further evaluate the potential application of our method in clinical practice, we recruited 36 gastric cancer cases, 29 patients with gastritis, and 71 individuals without gastric lesions, all of whom underwent endoscopy for diagnostic confirmation. Additionally, all samples in the Validation cohort had hotspot mutation results available for evaluation. Cancer diagnoses were confirmed for all patient participants, with staging determined according to the American Joint Committee on Cancer and International Union for Cancer Control (version VIII) guidelines. Healthy participants had no prior cancer history at enrollment, and their cancer-free status was verified by gastroscopy with a 12-month follow-up. Recruitment took place across multiple sites, including the University Medical Center, the tertiary hospital at Ho Chi Minh City, Vietnam, from May 2021 to April 2024.

The study methodologies adhered to the principles outlined in the Declaration of Helsinki and received approval from the Institutional Review Board, University Medical Center at Ho Chi Minh city and the Medical Genetics Institute in Ho Chi Minh City, Vietnam. Informed written consent was obtained from each participant, and all cancer patients were treatment-naïve at the time of blood sample collection.

Isolation of cfDNA and genomic DNA

Each participant provided 10 mL of peripheral blood, collected in a Cell-Free DNA BCT tube (Streck, USA). The isolation of cfDNA and genomic DNA (gDNA) is detailed in the Supplementary Materials.

To minimize technical variation and ensure consistency across runs, all blood samples were collected in Streck tubes, and plasma was isolated within 24 hours of blood draw. Plasma was stored at –80 °C for a maximum of 48 hours prior to DNA extraction. Cell-free DNA isolation and library preparation were performed centrally, with gastric cancer and control samples randomized across batches. Only samples with at least 2 ng of cfDNA input and that passed sequencing QC thresholds (TM depth coverage >10; percentage of mapped reads >70%) were included in downstream analyses.

Multimodal cfDNA assay

cfDNA samples were analyzed using a multi-feature ctDNA assay [6] to assess methylation across 450 targeted regions, genome-wide methylation, copy number variations, fragment length, and DNA end motifs. The workflow includes three main steps: (1) cfDNA isolated from blood is bisulfite-converted and adapter-ligated to create a whole-genome bisulfite library; (2) targeted enrichment via hybrid capture isolates 450 cancer-specific regions, while a re-hybridization step preserves genome-wide data. The target capture and whole-genome fractions are then sequenced on the DNBSEQ-G400 platform (MGI Tech, China) at ~52X and 0.55X read depths, respectively, yielding 100-bp paired-end reads. Sequencing data are demultiplexed with bcl2fastq (Illumina, USA) and quality-checked using FastQC v.0.11.9 and MultiQC v.1.12; (3) Pre-processed data yield four cfDNA feature sets—targeted methylation (TM), genome-wide methylation (GWM), fragment lengths, and end motifs (EM)—for machine learning model prediction. The detailed process is described in the Supplementary Materials.

To construct a binary classification machine learning model, five feature types—TMD, GWM, CNA, EM, and FLEN—from samples in the discovery cohort were used as input data. To

mitigate the effects of high dimensionality and retain pattern-based differences between cancer patients and healthy controls, we applied a minimal and inclusive approach to feature selection. All features significantly different between groups (FDR ≤ 0.05) were used as input for machine learning models, which capture non-linear interactions through regularization and cross-validation. Each feature type was modeled using four machine learning algorithms: Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGB). Models were trained and hyperparameters optimized with the caret package (version 6.0-94) using 10-fold cross-validation for robustness. Hyperparameter tuning was tailored to each algorithm: LR was tested with 'l1', 'l2', and 'none' penalties; RF used mtry values of 2, 3, and 4, with ntree set to 500; SVM employed various C and sigma values with a radial kernel; and XGB was configured for interaction depth, number of trees, shrinkage, and minimum observations per node. To enhance evaluation accuracy, we implemented nested 10-fold cross-validation, dividing data into 5 outer folds for testing and training, with an inner 5-fold cross-validation loop for hyperparameter tuning.

The best model, based on average ROC-AUC scores from the nested CV, was selected for further analysis. Additionally, we explored a combinatory strategy using a single data frame with all six features (ensemble model), following the same hyperparameter tuning and feature selection process.

Hotspot mutations - Amplicon-based sequencing

We analyzed somatic mutations among 94 GC patients in Vietnam and the COSMIC database, creating a panel of 535 hotspot. Hotspot mutations in cfDNA were detected by amplicon-based sequencing. Detailed methods are provided in the Supplementary Materials.

Statistical analysis

The Wilcoxon Rank Sum test or the t-test was employed to identify statistically significant differences between cancer and control features. The t-test was applied when the features adhered to a normal distribution; otherwise, the Wilcoxon Rank Sum test was used. The Kolmogorov-Smirnov test was utilized to determine if two random samples had the same statistical distribution. The statistical analysis is decribed in the Supplementary Materials.

RESULTS

Clinical characteristics of cancer and healthy participants

The study included a total of 352 participants, organized into two cohorts: a discovery cohort of 229 individuals (110 patients with GC and 119 healthy controls, **Figure 1**) and a validation cohort of 123 individuals (58 GC patients and 65 healthy controls, **Figure 1**). In the discovery cohort, the majority of GC patients were male (64.5%, **Table 1**), a proportion slightly higher than that of the healthy controls (58.0%, **Table 1**). However, gender distribution between the two groups did not differ significantly (Chi-Square test, $p = 0.309$, **Table 1**). Similarly, in the validation cohort, the male distribution was 56.9% among GC patients and 50.8% among healthy controls, with no statistically significant difference (Chi-Square test, $p = 0.496$, **Table 1**).

Age distribution revealed a significant disparity between GC patients and healthy controls in both cohorts. In the discovery cohort, the median age of GC patients was 63 years, significantly older than the median age of 52 years in healthy controls (Mann-Whitney test, $p < 0.0001$, **Table 1**). This trend was consistent in the validation cohort, where GC patients had a median age of 59 years versus 45 years for healthy controls (Mann-Whitney test, $p < 0.0001$, **Table 1**).

In terms of disease staging, early-stage cases (Stage 0-stage II) comprised 50.9% of the discovery cohort (with 35.5% in stages 0-I and 15.5% in stage II, **Table 1**) and 46.6% of the validation cohort (27.6% in stages 0-I and 19.0% in stage II, **Table 1**). Late-stage cases (Stage IIIA) accounted for 40.0% of the discovery cohort and a slightly higher proportion, 53.4%, in the validation cohort. Staging information was unavailable for 9.1% of patients in the discovery cohort; however, these patients were confirmed by specialized clinicians to have non-metastatic tumors.

Differential Methylation Analysis between GC Patients and Controls

Abnormal DNA methylation is a critical epigenetic signature in gastric tumorigenesis [11]. Previous research has primarily focused on assessing methylation changes in specific genomic regions involved in the regulation of gene expression [12]. In this study, we utilized the previously described workflow [6], which integrates bisulfite shallow genome-wide sequencing with deep target sequencing, to provide a comprehensive methylation profile of cfDNA in GC patients compared to healthy individuals. This approach enabled us to analyze methylation patterns both at specific genomic sites and across the entire genome.

For the target sequencing component, we profiled methylation changes across 450 selected genomic regions, chosen for their critical roles in the transcriptional regulation of cancer-associated genes. Among these, we identified 42 differentially methylated regions (DMRs) (Wilcoxon rank-sum test, Benjamini-Hochberg adjusted p-values <0.05, **Figure 2A** and **Table S2**) when comparing GC patients to health controls. Notably, several regions associated with multiple genes, including *CALN1*, *IKZF1*, *IFFO1*, *HMX1_CPZ*, *RASSF2A*, and *FGF5*, exhibited more than a two-fold increase in methylation in GC patients (**Figure 2B**).

In addition to site-specific DMRs, genome-wide hypomethylation is a significant and widespread epigenetic alteration associated with various cancers [6, 13, 14]. To investigate genome-wide methylation changes in GC patients, we mapped bisulfite sequencing reads from the whole-genome fraction to the human genome, which was divided into 1 Mb bins, resulting in a total of 2,734 bins across the genome (**Table S3**). The comparative analysis of methylation ratios across the 2,734 bins revealed substantial differences between GC patients and healthy controls. GC samples exhibited a broader distribution of methylation levels and significantly lower median methylation ratios compared to healthy individuals (Kolmogorov-Smimov test, $p < 0.0001$, **Figure 2C**). A chromosome-wide analysis identified 2,322 bins (81.6%) across all 22 chromosomes that were particularly susceptible to hypomethylation in the cancer samples (Wilcoxon rank sum test, Benjamini-Hochberg adjusted p-value <0.05, **Figure 2D**). Overall, this comprehensive methylation analysis of plasma cfDNA reveals distinct patterns that effectively differentiate GC patients from healthy individuals, highlighting the potential of cfDNA methylation profiling as a biomarker for GC.

Distinct DNA Copy Number Aberrations in GC Plasma cfDNA (Table S4)

In addition to genomic hypomethylation profile, DNA copy number aberration (CNA) is a critical hallmark of cancer, frequently associated with chromosomal instability, leading to gains or losses of chromosomal segments, or even entire chromosomes [15, 16]. Copy number aberrations are increasingly drawing interest in the detection of GC [17]. In this study, we compared CNA profiles in cfDNA between GC patients and healthy controls. We identified 571 bins (21%) with significant gains and 186 bins (7%) with significant CNA losses (Wilcoxon rank sum test, Benjamini-Hochberg adjusted p-value <0.05, **Figure 3A**). Notably, CNA gains were particularly prominent on chromosomes 17, 19, 20 and 22, while CNA losses were more frequent

on chromosomes 3 and 4 (**Figure 3B**). These findings provide new insights into the chromosomal alterations present in the cfDNA of GC patients, highlighting potential biomarkers for early detection.

Fragment Length Variability in GC Compared to Control Samples (Table S5)

Several studies have shown that cfDNA fragmentation patterns are non-random and influenced by apoptosis-mediated caspase activity, with ctDNA fragments generally shorter than non-cancer cfDNA [13, 18]. However, the fragment length profile of GC-derived cfDNA remains unclear. To address this, we conducted a comparative analysis of DNA fragment lengths between GC patients and healthy individuals. To validate our analytical approach, we included cfDNA fragment length data from 160 liver cancer patients who were enrolled in our previous study [19]. Consistent with previous findings, liver cancer patients exhibited a higher frequency of short DNA fragments (<150 bp) compared to healthy individuals (**Figure 3C**). In contrast, GC patients did not show the same enrichment of short fragments (**Figure 3C**). However, they did display a significant enrichment of fragments in the 150-170 bp range compared to healthy individuals (Wilcoxon rank sum test, Benjamini-Hochberg adjusted p-value <0.05, **Figure 3C** and **Figure 3D**). This suggests that cfDNA fragment length profiles are cancer type-specific and highlights the potential of cfDNA fragment size between 150 and 170 bp as a biomarker for GC detection.

Identification of Distinct End Motifs in cfDNA of GC (Table S6)

Differences in fragment length could be associated with variations in DNA motifs at the ends of each fragment, potentially arising from differential cleavage patterns in cancer cells compared to normal cells during apoptosis [20, 21]. We analyzed the differential presence of 256 possible 4-mer end motifs (EMs) in GC samples versus controls (Wilcoxon rank sum test, Benjamini-

Hochberg adjusted p-value <0.05). This revealed 87 motifs with significantly increased frequencies and 71 with decreased frequencies in GC samples (**Figure 4A**). The majority of motifs beginning with thymine are enriched in the cfDNA of GC patients, while the frequencies of motifs beginning with adenine are reduced (**Figure 4A**). Among the significant elemental motifs (EMs), the top five with the most significant frequency increases were CCGA, CCGC, CCGG, GCGC, and GCGG (**Figure 4B**), while the motifs with the most significant decreases in cancer patients compared to controls were CCCT, CCTT, GCCT, TACT, and TCTT (**Figure 4C**). These EM frequency shifts may serve as valuable markers for identifying ctDNA in GC.

Multimodal Analysis of cfDNA Features Could Improve GC Detection

The identification of significant signatures in both the targeted and genome-wide fractions of the discovery cohort led to the development of a multi-feature classification model to distinguish GC patients from healthy individuals. We constructed five feature datasets: Target Region Methylation (TM), Genome-wide Methylation (GWM), Copy Number Aberration (CNA), Fragment Length Distribution (FLEN), and End Motif (EM) (**Figure 5A**). We performed unsupervised clustering of the selected features using UMAP to assess potential batch effects across all samples, which were processed in 92 sequencing runs. The resulting UMAP plots showed gastric cancer and healthy control samples broadly intermixed across runs, with no evidence of batch-specific clustering (**Figure S2**). These results indicate that the selected features are not affected by technical variation in sample collection or processing.

Each dataset then underwent nested cross-validation using algorithms such as XGBoost (XGB), Logistic Regression (LG), Random Forest (RF), and Support Vector Machines (SVM). This nested cross-validation included five outer iterations, each with a five-fold inner cross-validation phase for hyperparameter tuning and feature selection, ensuring robust performance metrics. We

compared the individual feature models with an ensemble model that integrated all five features. The performance of the single-feature models varied, with the GWM and EM-based models generally outperforming the others during the discovery phase, achieving an AUC of 0.85 (95% CI: 0.80–0.90) (**Figure 5B**). An ensemble model, constructed by combining the five best-performing single-feature models using logistic regression, further improved classification performance, achieving an AUC of 0.90 (95% CI: 0.86–0.94).

Given the importance of high specificity in large-scale cancer screening to reduce false positives and minimize psychological impact, we established a cut-off value for each model ensuring a minimum specificity threshold of 95%. The ensemble model consistently outperformed the single-feature models, achieving a sensitivity of 0.65 (95%CI: 0.51-0.77) at a specificity of 96% (**Figure 5D**). To further assess the robustness of the ensemble model, we tested its performance on an independent validation cohort of 58 GC patients and 65 healthy individuals. Consistent with the discovery cohort, the ensemble model demonstrated high accuracy in detecting GC in validation cohort, achieving an AUC of 0.87 (95% CI: 0.80–0.93) (**Figure 5C**), with a sensitivity of 0.71 and specificity of 92% (**Figure 5E**).

To further evaluate the potential application of our method in clinical practice, we additionally recruited 36 gastric cancer cases, 29 patients with gastritis, and 71 individuals without gastric lesions, all of whom underwent endoscopy for diagnostic confirmation. In this external validation cohort, our model achieved a sensitivity of 72.2%, with specificities of 94.4% in individuals without gastric lesions and 89.7% in patients with benign gastric conditions, yielding an overall specificity of 93.0% (**Figure S3**). These results demonstrate the consistent performance of our assay in an independent cohort.

Assessing the Benefit of Hotspot Mutations in Multimodal Detection

Cancer-specific mutations in plasma cfDNA have emerged as promising biomarkers for cancer detection [7]. To investigate this potential in GC, we developed a deep amplicon-based sequencing workflow to profile hotspot mutations in the cfDNA of GC patients. This workflow utilized a panel of 535 hotspot mutations curated from the COSMIC database, supplemented with somatic mutations identified in 94 Vietnamese GC patients. To differentiate cancer-specific mutations from clonal hematopoiesis-associated variants, we also performed deep sequencing on matched gDNA from white blood cells (WBCs). Among the 58 cancer patients in the validation cohort, 18 (31.0%) GC patients had at least one hotspot mutation in plasma cfDNA, while no mutations were detected in 65 healthy individuals (**Figure 6A**). Most hotspot mutations originated from *TP53*, detected in 10 (17%) of all GC patients (**Figure 6A**).

We next investigated whether incorporating hotspot mutation analysis into our multimodal framework based on methylation and fragmentomic features could improve the sensitivity of GC detection. Notably, 15 GC patients (25.9%) were concordantly detected by both hotspot mutation analysis and the ensemble model, while 3 (5.2%%) and 26 (44.8%) patients were uniquely detected by hotspot mutation analysis or the ensemble model, respectively (**Figure 6B**). This discordance highlights the potential advantage of combining these approaches to enhance detection rates. For early-stage GC (Stages 0-II), the integrated method achieved a sensitivity of 74.1%, compared to 14.8% and 70.4% using hotspot mutation analysis or the ensemble model alone, respectively (**Figure 6C**). In advanced cases (Stage III), the combined approach reached a sensitivity of 77.4%, compared to 45.2% and 71.0% for hotspot mutation analysis or the ensemble model alone (**Figure 6C**). These findings suggest that incorporating hotspot mutation analysis using an OR logic into the multimodal framework could significantly enhance the ensemble model's overall performance in detecting GC.

We next evaluated the performance of the ensemble model in comparison to our previously developed SPOT-MAS model, which is based on features identified through pairwise comparisons between healthy controls and multiple cancer types. Using the same validation cohort, we excluded samples previously analyzed in the original publication, resulting in 23 GC samples and 65 healthy controls. The ensemble model outperformed the SPOT-MAS model, demonstrating significantly higher sensitivity for GC detection across all stages (73.9% vs. 47.8%) and particularly for early-stage cancers (Stages I-II) where sensitivity was 72.7% compared to 36.4% for SPOT-MAS (**Figure 6D**). Both models maintained similar specificity, with the ensemble model achieving 92.3% and the SPOT-MAS model achieving 90.8% (**Figure 6D**). These results underscore that training the ensemble model with GC-specific cfDNA signatures substantially enhances its performance over models trained on generalized, multi-cancer features.

DISCUSSION

GC presents a substantial public health challenge, underscoring the critical need for effective early detection strategies, particularly among high-risk populations [22]. However, an established standard of care for GC screening, especially in asymptomatic individuals, remains lacking. Currently, endoscopy—though considered the gold standard for GC detection—is invasive, resource-intensive, and is not commonly used as a screening tool outside of regions with high prevalence [23]. In this study, we introduce a minimally invasive ctDNA-based assay developed to enhance the early detection of GC.

Our study introduces a novel pair-wise comparison analysis, specifically designed to identify GC-specific features within cfDNA. Among the 42 differentially methylated regions (DMRs) identified in our study, several, including *CALN1*, *IKZF1*, and *IFFO1*, have been previously

associated with gastric carcinogenesis [24-26]. Notably, we also identified additional DMRs specific to GC, such as *TMEM260_PELI2*, a region implicated in cell signaling pathways that may contribute to tumorigenesis and disease progression [27]. The identification of *TMEM260* in this expanded analysis highlights its potential as a novel biomarker for GC, thereby enriching our understanding of genes involved in GC pathogenesis [6]. We observed a pattern of genome-wide hypomethylation in cfDNA from GC patients relative to healthy individuals [28]. This finding aligns with prior studies indicating that global hypomethylation is a hallmark of tumorigenesis, often linked to genomic instability in cancer [29-31]. The validation of this pattern in our study strengthens the association between hypomethylation and early GC development, supporting the hypothesis that epigenetic changes play a pivotal role in cancer initiation and progression.

Our study also assessed CNA profiles in GC compared to those reported in the pan-cancer analysis from our previous model [6]. While considerable overlap in CNAs was observed across cancer types, GC exhibited unique CNA patterns, particularly with notable CNA gains on chromosomes 19 and 22. These alterations have been consistently documented in other GC studies, underscoring their potential relevance as disease-specific markers [32]. Another key finding was the distinct fragment length profiles observed in ctDNA from GC patients compared to patients with colorectal cancer (CRC) [33] and hepatocellular carcinoma (HCC) [14], as noted in previous research. While ctDNA fragments from CRC and HCC patients generally showed enrichment in shorter fragments (<150 bp) [34], GC patients exhibited an enrichment in the 150–170 bp range, which may reflect unique nucleosome footprints or differences in gene expression profiles across cancer types. Consistent with our previous findings, we observed that combining multiple features, such as methylation and fragmentomics, enhanced classification performance compared to using either feature alone [6]. This integrated approach provided a more robust

model for detecting early-stage GC, in line with existing literature emphasizing the complementary nature of these features in capturing the complexity of tumorigenesis [14].

In the external validation cohort, a slight reduction in specificity was observed in the gastric benign group. This likely reflects a limitation of the training strategy, as the discovery cohort did not include benign gastric conditions. Consequently, the model was not explicitly trained to distinguish cancer from non-malignant but clinically similar presentations. Benign lesions often share cfDNA fragmentation and methylation features with cancer, which can increase the likelihood of misclassification [35, 36]. Incorporating benign cases into future training cohorts will therefore be critical to further improving model robustness and reducing false positives in real-world clinical settings.

To investigate the importance of features selected by our model, we conducted SHAP (SHapley Additive exPlanations) analysis (**Figure S4**). Among the top-ranked features, the end motif CTGA emerged as the most important, showing significant enrichment in gastric cancer patients (**Figure 4A**). This observation is consistent with a recent study reporting that end motifs beginning with "C" were significantly enriched in early-stage lung cancer [37], potentially linked to dysregulated *DNASE1L3* activity previously described in cancer [38]. SHAP analysis also highlighted abnormal hypomethylation events on chromosomes 15 and 2. Hypomethylation is a well-established driver of genomic instability, a hallmark of carcinogenesis. Prior studies have documented the vulnerability of chromosome 15 [39] and chromosome 2 [40] in gastric tumorigenesis. Genomic instability was further reflected at the level of abnormal CNAs. Our SHAP analysis identified the most significant CNA feature on chromosome 18, with three high-impact bins selected. This is consistent with recurrent chromosome 18q loss in gastric cancer, a poor prognostic event associated with tumor suppressor inactivation and oncogene activation

[41, 42]. Similarly, recurrent chromosome 4q deletions have been reported in both intestinal and diffuse histological subtypes [43], while chromosome 8q gains are among the most frequent oncogenic CNAs across gastrointestinal cancers [40]. Finally, the promoter region of *SAMD4A* emerged as a key methylation feature in our SHAP analysis. Notably, hypermethylation of this promoter has been reported as a biomarker for predicting clinical outcomes in gastric cancer [44]. Collectively, these associations between SHAP-derived feature importance and established biological functions in gastric cancer indicate that the features selected by our model make biologically meaningful contributions to the classification performance of our multi-feature assay.

In addition, evaluating model reproducibility and determining the limit of detection (LOD) are essential to ensure that the classifier performs consistently across datasets and experimental conditions, and to define the lowest tumor fraction at which the model can reliably distinguish cancer from non-cancer samples. To assess reproducibility, we analyzed six pooled plasma samples (three gastric cancer and three healthy controls) that were divided into three independent batches (**Figure S5A**). We observed strong correlations for all analyzed features across batches (Pearson test, p-value < 0.05), with comparable prediction scores for both healthy and gastric cancer samples (**Figure S5B–C**). These findings demonstrate highly consistent assay performance across independent runs.

For LOD determination, we estimated ctDNA tumor fraction (TF) using the ichorCNA algorithm [45], which is designed for ultra-low coverage WGS data and can robustly estimate TF without requiring prior tumor genotyping. To generate material for LOD testing, we first pooled cfDNA from 63 healthy individuals and 6 gastric cancer patients. The TF of these pooled samples was estimated using ichorCNA, after which we created spike-in mixtures by blending cancer cfDNA

pools into healthy cfDNA pools at predefined TFs of 0.5%, 1%, 5%, 15%, 25%, 50%, and 100% (**Figure S6A**). These mixtures were then analyzed using our multi-feature assay. By fitting a detection curve of sensitivity versus TF at a fixed specificity of 96% (cut-off defined in the discovery phase), we determined that the assay achieves 50% sensitivity at a tumor fraction of approximately 0.052 (LOD50, 95%CI 0.034-0.070; **Figure S6B**). This demonstrates that our assay can reliably detect cancer-derived cfDNA at low tumor fractions, providing a quantitative benchmark for its analytical sensitivity in early cancer detection.

Our methylation- and fragmentomics-based model outperformed hotspot mutation-based approaches for early-stage GC detection (stage 0-II, **Figure 6C**), suggesting that methylation and fragmentomic changes are more prevalent during early tumorigenesis. This observation aligns with the known low mutation burden in GC, which limits the effectiveness of mutation-based detection strategies [46]. Notably, incorporating hotspot mutations into our ensemble model using OR logic further improved detection sensitivity, particularly for patients with late-stage (Stage III) GC, while maintaining comparable specificity (**Figure 6C**). These findings indicate that epigenetic and genetic alterations may serve as complementary, non-redundant biomarkers in advanced disease stages. Additionally, when comparing our ensemble model to the previously developed SPOT-MAS model, which utilizes features derived from multiple cancer types, the ensemble model demonstrated superior sensitivity (73.9% vs. 47.8%, **Figure 6D**) across all stages, with the advantage especially pronounced in early-stage tumors (72.7% vs. 36.4%, **Figure 6D**). These results highlight the value of training detection models with cancer type-specific cfDNA signatures to optimize performance.

Compared to multi-cancer early detection (MCED) assays that include gastric cancer – such as CancerSEEK [7] (which combines protein biomarkers and hotspot mutations), Galleri (which

relies on methylation), and SPOT-MAS [6] (which integrates methylation and fragmentomics) – our assay demonstrated higher sensitivity but slightly lower specificity (**Table S7**). When benchmarked against assays specifically designed for gastric cancer detection, our test achieved comparable specificity (92%) while offering substantially higher sensitivity for early-stage gastric cancer than the methylation-only assay developed by Ren et al. [8] (75.9% vs. 44–59%, **Table S7**), underscoring the advantage of a multi-feature approach. In comparison with the multi-feature assay reported by Yu et al. [47], our assay achieved higher specificity (92.3% vs. 89.8–91.5%, **Table S7**) but somewhat lower sensitivity (75.9% vs. 87.2–91.8%, **Table S7**). This difference may, in part, reflect sequencing depth: our assay employs shallow whole-genome sequencing at 0.55× coverage to reduce cost and data complexity, making it more practical for large-scale screening. By contrast, Yu et al.'s assay was performed at 8× coverage, which may enhance sensitivity but introduces challenges in terms of cost-effectiveness and scalability [47]. We have acknowledged in the revised manuscript that future large-scale studies are warranted to systematically evaluate the trade-offs between sensitivity, sequencing depth, cost, and clinical applicability in real-world population screening programs.

Nonetheless, certain limitations should be considered. First, differences in clinical features between healthy individuals and GC patients could potentially confound ctDNA characteristics, such as methylation and fragment length profiles. These differences may reflect demographic variables like age and gender rather than cancer-specific alterations. We accounted for this by assessing the impact of gender and age on the performance of our model. Specifically, we examined whether differences in age and gender distributions in our cohorts could confound model predictions. Specifically, we examined whether differences in age and gender distributions in our cohorts could confound model predictions. Specifically, we compared model-generated

probabilities of having gastric cancer between participants aged ≤59 years and >59 years (59 being the median age of the discovery cohort). No significant difference in model scores was observed between the two age groups (**Figure S1**). Similarly, there were no statistically significant differences in prediction scores between male and female participants in either the gastric cancer or control groups (**Figure S1**). These results suggest that our model performs robustly across age and gender subgroups and is not significantly influenced by these demographic variables. Additionally, the retrospective nature of our study may introduce biases associated with sample collection, storage, and processing. A prospective study would provide a more rigorous assessment of the model's real-world clinical performance, helping to mitigate potential biases. Another limitation is the relatively low depth of bisulfite sequencing, which may reduce sensitivity to detect genome-wide methylation changes. However, the reduced sequencing depth provides practical advantages, including lower costs and faster processing times, which are critical for clinical applications. Finally, our study did not address the capacity of model to distinguish between GC and benign gastric lesions, such as gastritis or polyps. This distinction is essential to reduce false positives and improve the model's clinical applicability. Therefore, future studies should focus on prospective validation in larger cohorts, including patients with benign gastric conditions, to rigorously assess the clinical utility of this multimodal cfDNA-based assay.

CONCLUSIONS

This study underscores the potential of a novel multimodal cfDNA assay that integrates methylation and fragmentomic profiling for the early detection of GC. The assay demonstrated robust diagnostic performance, significantly enhancing sensitivity for early-stage detection, particularly when combined with hotspot mutation analysis. These findings suggest that this

innovative approach could improve early diagnosis and ultimately enhance patient outcomes in GC.

Declarations

Ethics approval and consent to participate

This study received approval from the Institutional Review Board, University Medical Center at Ho Chi Minh city, as well as the Medical Genetics Institute in Ho Chi Minh City, Vietnam (ID: 79/GCN-HĐĐĐ). Informed written consent was obtained from each participant in accordance with the Declaration of Helsinki.

Consent for publication

Not applicable

Availability of data and materials

The datasets generated during the current study are available in the NCBI SRA repository (PRJNA1308326), [https://www.ncbi.nlm.nih.gov/sra/PRJNA1308326]. The complete code—including preprocessing, feature extraction, and model training— is available in the GitHub repository, [https://github.com/huynhkl953/Gastric-Paper].

Competing interests

The authors, including LST, HG, MDP, and HNN, hold equity in Gene Solutions. HG, MDP, and LST are listed as inventors on the patent application (USPTO 17930705). We confirm that this does not impact on our compliance with the journal's policies regarding data and material sharing.

Funding

This work was supported by Gene Solutions.

Authors' contribution:

DLV, LAKH, CVTN, THHN, DHV, TTVV, TTN1, ATN, HGV, THN perfomed formal analysis.

DLV, TND, VHN, QDT, QTD, VTAN, LMQH, TPDH, TNDD, PTNN, KTN, VCH, TLL, THNN, NHT, TST, TXJ, TLV, THTN, TTT, MHT, NCT, TCN, TTN2, BTL, VPT, TVP, TNTN, HAC performed patient consultancy and screening.

LAKH, CVTN, THHN, DHV, TTVV, THN performed data curation.

LNT, MDP, HG, HNN LST performed the methodology.

MDP, HG, HNN, LST performed conceptualization.

LAKH, CVTN, GTHN, LST performed writing-original draft.

LAKH, CVTN, GTHN, LST performed writing-review and editing.

TTN1 corresponding to Thi Tu Nguyen and TTN2 corresponding to Thi Truc Nguyen

Acknowledgments:

We extend our gratitude to all participants for their involvement in this study, as well as to the clinics and hospitals that supported patient consultations and facilitated sample collection.

REFERENCE

1. Shin, W.S., et al., *Updated Epidemiology of Gastric Cancer in Asia: Decreased Incidence but Still a Big Challenge.* Cancers (Basel), 2023. **15**(9).
2. Bray, F., et al., *Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA: A Cancer Journal for Clinicians, 2024. **74**(3): p. 229-263.
3. Mai, T.T.T., et al., *Characteristics and health problems of cancer patients admitted to palliative care service at the Oncology Hospital in Ho Chi Minh City, Vietnam: a cross-sectional study.* MedPharmRes, 2024. **8**(2): p. 90-103.

4. In, H., et al., *Pepsinogen and Helicobacter pylori: Serum biomarkers for gastric cancer risk in a diverse United States population.* Surgical Oncology Insight, 2024. **1**(3).
5. De Rubis, G., S. Rajeev Krishnan, and M. Bebawy, *Liquid Biopsies in Cancer Diagnosis, Monitoring, and Prognosis.* Trends Pharmacol Sci, 2019. **40**(3): p. 172-186.
6. Nguyen, V.T.C., et al., *Multimodal analysis of methylomics and fragmentomics in plasma cell-free DNA for multi-cancer early detection and localization.* eLife, 2023. **12**: p. RP89083.
7. Cohen, J.D., et al., *Detection and localization of surgically resectable cancers with a multi-analyte blood test.* Science, 2018. **359**(6378): p. 926-930.
8. Ren, J., et al., *Genome-Scale Methylation Analysis of Circulating Cell-Free DNA in Gastric Cancer Patients.* Clin Chem, 2022. **68**(2): p. 354-364.
9. Liu, M.C., et al., *Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA.* Ann Oncol, 2020. **31**(6): p. 745-759.
10. Bettegowda, C., et al., *Detection of circulating tumor DNA in early- and late-stage human malignancies.* Sci Transl Med, 2014. **6**(224): p. 224ra24.
11. Li, J.H., et al., *Clinical applications and perspectives of circulating tumor DNA in gastric cancer.* Cancer Cell Int, 2024. **24**(1): p. 13.
12. Buitrago, D., et al., *Impact of DNA methylation on 3D genome structure.* Nature Communications, 2021. **12**(1): p. 3243.
13. Nguyen, V.-C., et al., *Fragment length profiles of cancer mutations enhance detection of circulating tumor DNA in patients with early-stage hepatocellular carcinoma.* BMC Cancer, 2023. **23**(1): p. 233.
14. Phan, T.H., et al., *Circulating DNA Methylation Profile Improves the Accuracy of Serum Biomarkers for the Detection Of Nonmetastatic Hepatocellular Carcinoma.* Future Oncology, 2022. **18**(39): p. 4399-4413.
15. Dereli-Öz, A., G. Versini, and T.D. Halazonetis, *Studies of genomic copy number changes in human cancers reveal signatures of DNA replication stress.* Mol Oncol, 2011. **5**(4): p. 308-14.
16. Baldacchino, S. and G. Grech, *Somatic copy number aberrations in metastatic patients: The promise of liquid biopsies.* Semin Cancer Biol, 2020. **60**: p. 302-310.
17. Keller, L., et al., *Clinical relevance of blood-based ctDNA analysis: mutation detection and beyond.* Br J Cancer, 2021. **124**(2): p. 345-358.
18. Cristiano, S., et al., *Genome-wide cell-free DNA fragmentation in patients with cancer.* Nature, 2019. **570**(7761): p. 385-389.
19. Pham, T.M.Q., et al., *Multimodal analysis of genome-wide methylation, copy number aberrations, and end motif signatures enhances detection of early-stage breast cancer.* Front Oncol, 2023. **13**: p. 1127086.
20. Jiang, P., et al., *Plasma DNA End-Motif Profiling as a Fragmentomic Marker in Cancer, Pregnancy, and Transplantation.* Cancer Discov, 2020. **10**(5): p. 664-673.
21. Jin, C., et al., *Characterization of fragment sizes, copy number aberrations and 4-mer end motifs in cell-free DNA of hepatocellular carcinoma for enhanced liquid biopsy-based cancer detection.* Mol Oncol, 2021. **15**(9): p. 2377-2389.
22. Morgan, E., et al., *The current and future incidence and mortality of gastric cancer in 185 countries, 2020-40: A population-based modelling study.* EClinicalMedicine, 2022. **47**: p. 101404.
23. Yoshida, N., et al., *Early gastric cancer detection in high-risk patients: a multicentre randomised controlled trial on the effect of second-generation narrow band imaging.* Gut, 2021. **70**(1): p. 67-75.
24. Li, H., et al., *Overexpression of lncRNA H19 enhances carcinogenesis and metastasis of gastric cancer.* Oncotarget, 2014. **5**(8): p. 2318-29.
25. Laven-Law, G., et al., *BCAT1, IKZF1 and SEPT9: methylated DNA biomarkers for detection of pan-gastrointestinal adenocarcinomas.* Biomarkers, 2024. **29**(4): p. 194-204.

26. Zhang, S., et al., *Discriminating Origin Tissues of Tumor Cell Lines by Methylation Signatures and Dys-Methylated Rules.* Front Bioeng Biotechnol, 2020. **8**: p. 507.
27. Polikarpova, A.V., et al., *CRISPR/Cas9-generated mouse model with humanizing single-base substitution in the Gnao1 for safety studies of RNA therapeutics.* Front Genome Ed, 2023. **5**: p. 1034720.
28. Chan, K.C., et al., *Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing.* Proc Natl Acad Sci U S A, 2013. **110**(47): p. 18761-8.
29. Hon, G.C., et al., *Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer.* Genome Res, 2012. **22**(2): p. 246-58.
30. Lei, Z.N., et al., *Signaling pathways and therapeutic interventions in gastric cancer.* Signal Transduct Target Ther, 2022. **7**(1): p. 358.
31. Zhang, S., et al., *Tumor initiation and early tumorigenesis: molecular mechanisms and interventional targets.* Signal Transduct Target Ther, 2024. **9**(1): p. 149.
32. Varis, A., et al., *DNA copy number changes in young gastric cancer patients with special reference to chromosome 19.* Br J Cancer, 2003. **88**(12): p. 1914-9.
33. Nguyen, H.T., et al., *Multimodal Analysis of ctDNA Methylation and Fragmentomic Profiles Enhances Detection of Nonmetastatic Colorectal Cancer.* Future Oncology, 2022. **18**(35): p. 3895-3912.
34. Guo, J., et al., *Quantitative characterization of tumor cell-free DNA shortening.* BMC Genomics, 2020. **21**(1): p. 473.
35. Gao, Q., et al., *Circulating cell-free DNA for cancer early detection.* Innovation (Camb), 2022. **3**(4): p. 100259.
36. Van, T.T.V., et al., *Multimodal analysis of cell-free DNA enhances differentiation of early-stage breast cancer from benign lesions and healthy individuals.* BMC Biol, 2025. **23**(1): p. 259.
37. Lee, T.-R., et al., *Integrating Plasma Cell-Free DNA Fragment End Motif and Size with Genomic Features Enables Lung Cancer Detection.* Cancer Research, 2025. **85**(9): p. 1696-1707.
38. Deng, Z., et al., *DNASE1L3 as a Prognostic Biomarker Associated with Immune Cell Infiltration in Cancer.* Onco Targets Ther, 2021. **14**: p. 2003-2017.
39. Kitayama, Y., et al., *Nonrandom chromosomal numerical abnormality predicting prognosis of gastric cancer: a retrospective study of 51 cases using pathology archives.* Lab Invest, 2003. **83**(9): p. 1311-20.
40. Kitayama, Y., H. Igarashi, and H. Sugimura, *Different Vulnerability among Chromosomes to Numerical Instability in Gastric Carcinogenesis: Stage-dependent Analysis by FISH with the Use of Microwave Irradiation1*. Clinical Cancer Research, 2000. **6**(8): p. 3139-3146.
41. Matthews, S., et al., *Variable Gene Copy Number in Cancer-Related Pathways Is Associated With Cancer Prevalence Across Mammals.* Mol Biol Evol, 2025. **42**(3).
42. Nemtsova, M.V., E.B. Kuznetsova, and I.V. Bure, *Chromosomal Instability in Gastric Cancer: Role in Tumor Development, Progression, and Therapy.* Int J Mol Sci, 2023. **24**(23).
43. Nishiu, M., et al., *Distinct pattern of gene expression in pyothorax-associated lymphoma (PAL), a lymphoma developing in long-standing inflammation.* Cancer Sci, 2004. **95**(10): p. 828-34.
44. Park, J.E., et al., *Diffsig: Associating Risk Factors with Mutational Signatures.* Cancer Epidemiol Biomarkers Prev, 2024. **33**(5): p. 721-730.
45. Adalsteinsson, V.A., et al., *Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors.* Nat Commun, 2017. **8**(1): p. 1324.
46. Kim, J., et al., *Tumor Mutational Burden Determined by Panel Sequencing Predicts Survival After Immunotherapy in Patients With Advanced Gastric Cancer.* Front Oncol, 2020. **10**: p. 314.
47. Yu, P., et al., *Multi-dimensional cell-free DNA-based liquid biopsy for sensitive early detection of gastric cancer.* Genome Medicine, 2024. **16**(1): p. 79.

Figures

Figure 1: Study design and workflow for sample recruitment and analysis.

Between May 2021 and July 2024, non-metastatic gastric cancer (GC) patients and control individuals were recruited from the University Medical Center, Ho Chi Minh City (168 GC patients and 184 controls). The discovery cohort, comprising 110 GC patients and 119 controls, underwent multimodal cfDNA analysis to assess fragmentomic and methylation profiles, which facilitated the development of machine learning models. A validation cohort of 58 GC patients and 65 controls was utilized to assess the performance of the model. Hotspot mutation panels were generated by analyzing the frequency of somatic mutations from the COSMIC database and an in-house cohort of 94 gastric cancer patients. Hotspot mutation profiles in cfDNA samples from the validation cohort were analyzed using amplicon-based sequencing

Figure 2: Analysis of methylation levels between controls and gastric cancer on both genome-wide and targeted scales.

(A) Volcano plot displaying the $\log_2$ fold-change of methylation levels versus the statistical significance ($-\log_{10}$ p-value) for specific genomic regions.

(B) Boxplot illustrating the comparison of methylation levels of selected genes (*CALN1*, *IKZF1*, *IFFO1*, *HMX1_CPZ*, *RASSF2A*, and *FGF5*) between control and gastric cancer groups (Mann-Whitney test, Mean ± SEM).

(C) Density plot comparing methylation ratio distributions between gastric cancer (red) and control (blue) groups, with statistical significance assessed using a Kolmogorov-Smirnov test.

(D) Point plot showing genome-wide methylation differences across chromosomes, with points color-coded by methylation status: hypermethylated (red, indicating increased methylation

relative to controls), hypomethylated (blue, indicating decreased methylation relative to controls), or no significant change (grey).

Figure 3: Analysis of Copy number aberrations (CNA) and Fragment length between control and gastric cancer.

(A) Scatter plot of CNA log2 fold-change across 22 chromosomes, indicating regions with CNA gain (red, regions with increased copy number relative to normal), CNA loss (blue, regions with decreased copy number relative to normal), and no significant change (grey).

(B) Bar plot showing the percentage of bins with CNA gain, CNA loss, and no significant change across each chromosome.

(C) Density plot comparing fragment length distributions across liver cancers (green, n = 160), gastric cancers (red, n = 110), and control (blue, n = 119) groups.

(D) Heatmap displaying $\log_2$ fold-change ($\log_2$FC) in fragment lengths, highlighting significant differences (t-test, p-value < 0.05) between gastric cancer and control groups. Red represents an increase in fragment length, while blue indicates a decrease.

Figure 4: End motif analysis between control and gastric cancer.

(A) Heatmap showing $\log_2$ fold-change ($\log_2$FC) of various end motifs (A, T, G and C) between control and gastric cancer groups. Red indicates enrichment, while blue indicates depletion in the gastric cancer group.

(B) Boxplot comparing the frequency of specific end motifs between control (blue) and gastric cancer (red) groups.

(C) Boxplot showing the frequency of additional end motifs between control (blue) and gastric cancer (red) groups, highlighting significant differences.

Figure 5: Workflow of Machine Learning Model for Classifying Gastric Cancer vs. Control Groups.

(A) Workflow diagram illustrating the construction and validation of classification models. The discovery cohort included 110 gastric cancer patients and 119 controls. Data preprocessing was performed on five feature sets: targeted methylation (TM) genome-wide methylation (GWM), copy number alterations (CNA), fragment length (FLEN), and end motifs (EM). Gradient Boosting Machine (GBM), Logistic Regression (LG), Random Forest (RF), and Support Vector Machine (SVM) models were trained using nested cross-validation, with hyperparameter tuning and feature selection incorporated. The final models were evaluated on the validation cohort (58 cancer patients, 65 controls), with the cutoff selected to ensure specificity >95%.

(B) ROC curves illustrating the performance of single feature based models and the ensemble model in the discovery cohort.

(C) ROC curves illustrating the performance of individual feature models and the ensemble model in the validation cohort.

(D) Bar plots comparing sensitivity and specificity across different models in the discovery cohort. (E) Bar plots comparing sensitivity and specificity across different models in the validation cohort.

Figure 6: Hotspot Mutation Analysis in the Validation Cohort.

(A) Heatmap displaying the detection of hotspot mutations and the ensemble model for 58 cancer patients and 65 healthy controls in the validation cohort. Each column represents the detection of a specific hotspot mutation or ensemble model (above the cutoff value), while each row

corresponds to an individual sample. Tumor stage and variant allele frequency (VAF) are depicted using color gradients for clarity

(B) Venn diagram showing the unique and overlapping detections between HOTSPOT and the ensemble model.

(C) Bar plot illustrating the accuracy of the HOTSPOT, Ensemble, and combined HOTSPOT and Ensemble across various stages (II/III, III), cases with unknown staging, all gastric cancer cases, and control groups.

(D) Bar plot comparing the accuracy of the SPOT-MAS and the Ensemble model across different stages (II/III, III), unknown staging, all gastric cancer cases, and control groups.

Supplement Figure 1: **Association between patients' gender and age and the performance of the ensemble model**

Supplement Figure 2: **UMAP analysis on selected features across sequencing runs for (A) gastric cancers patients and (B) healthy controls**

Supplement Figure 3: **Performance of external validation**

Supplement Figure 4: SHAP analysis on selected features

Supplement Figure 5: Reproducibility of features and ensemble model predictions across three independent batches.

Supplement Figure 6: **Limit of detection estimation from spike-in series of tumor cfDNA.** (A) Workflow for generating spike-in mixtures by blending pooled gastric cancer cfDNA into

healthy cfDNA pools at predefined tumor fractions (TFs; ichorCNA-estimated) of 0.5%, 1%, 5%, 15%, 25%, 50%, and 100%.

(B) Sigmoid curve illustrating the limit of detection (LOD) for gastric cancer. LOD values are defined as the TF level at which the probability of detecting a cancer signal reaches at least 50% while maintaining ≥96% specificity.

Tables

Table 1: Summary of clinical information of 204 gastric cancer patients and 284 non-cancer subjects (total N = 488)

Table S1: Clinical information of cancer patients and healthy participants

Table S2: List of 450 target regions.

Table S3: List of 2734 GWMD bins.

Table S4: List of 2691 CNA bins.

Table S5: List of 151 fragment lengths.

Table S6: List of 256 end motifs.

Table S7: Comparison of Studies on Early Gastric Cancer Detection

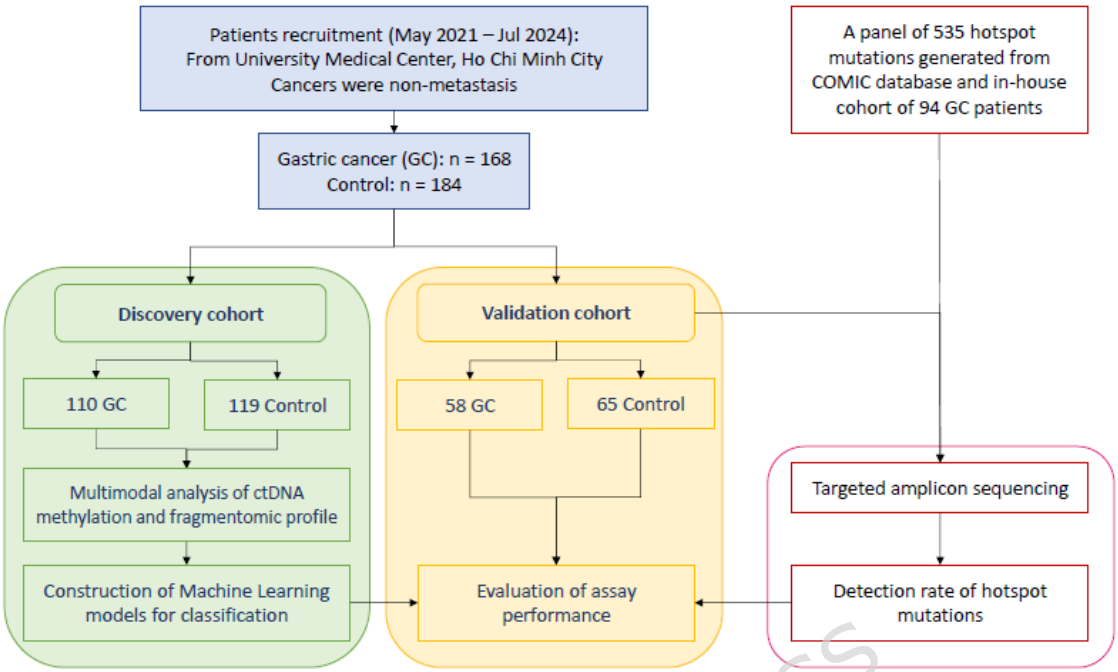

Figure 1

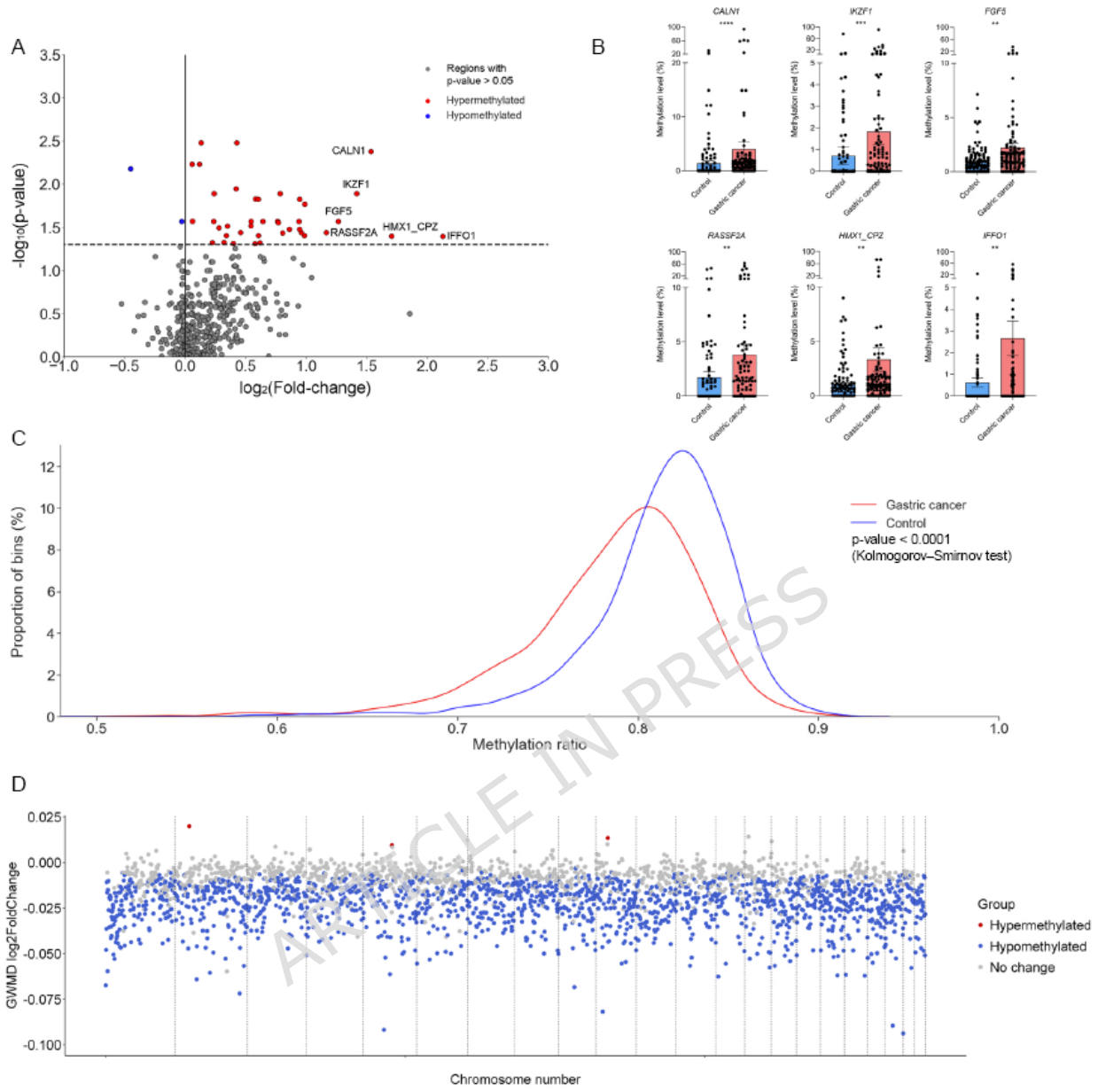

Figure 2

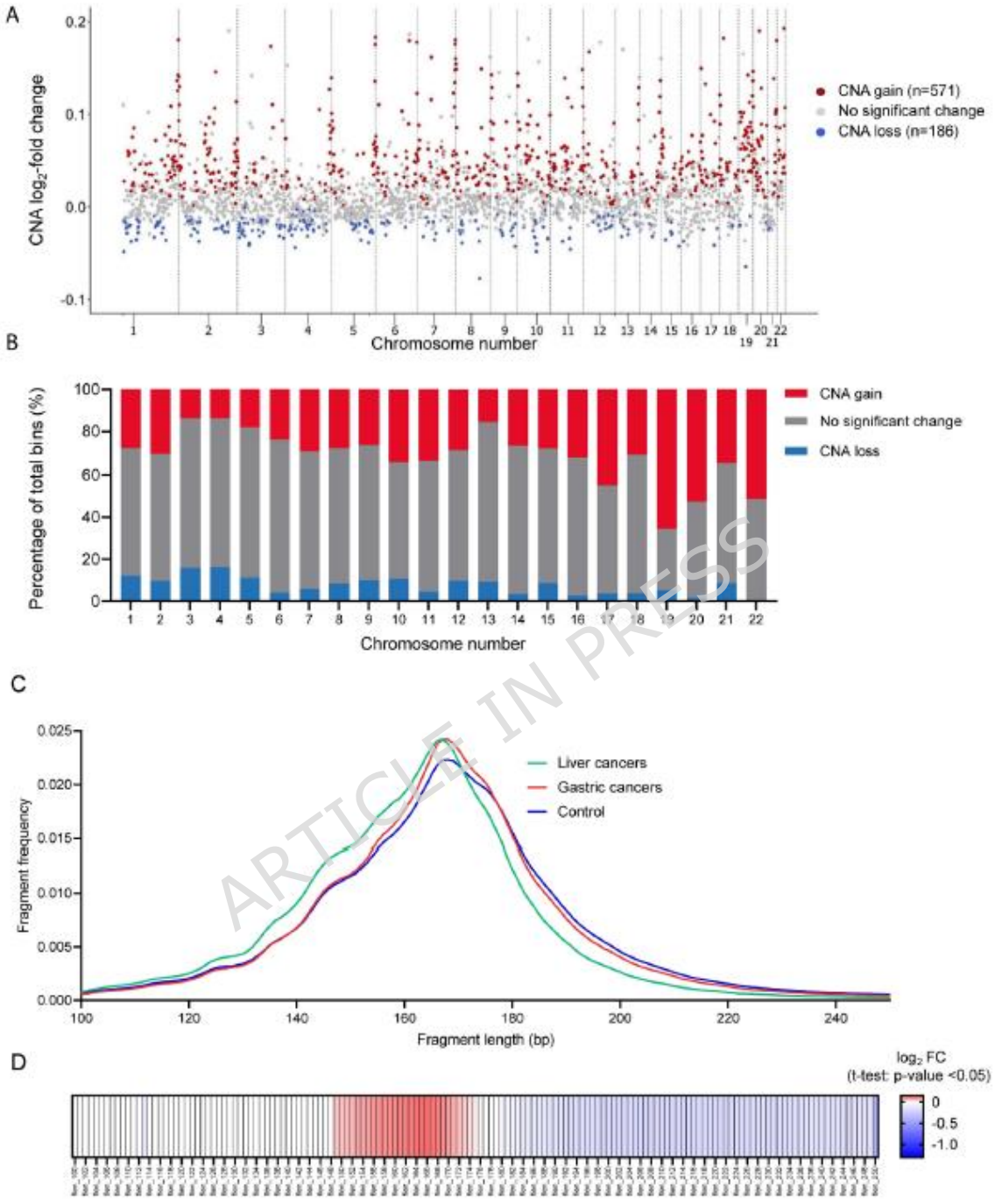

Figure 3

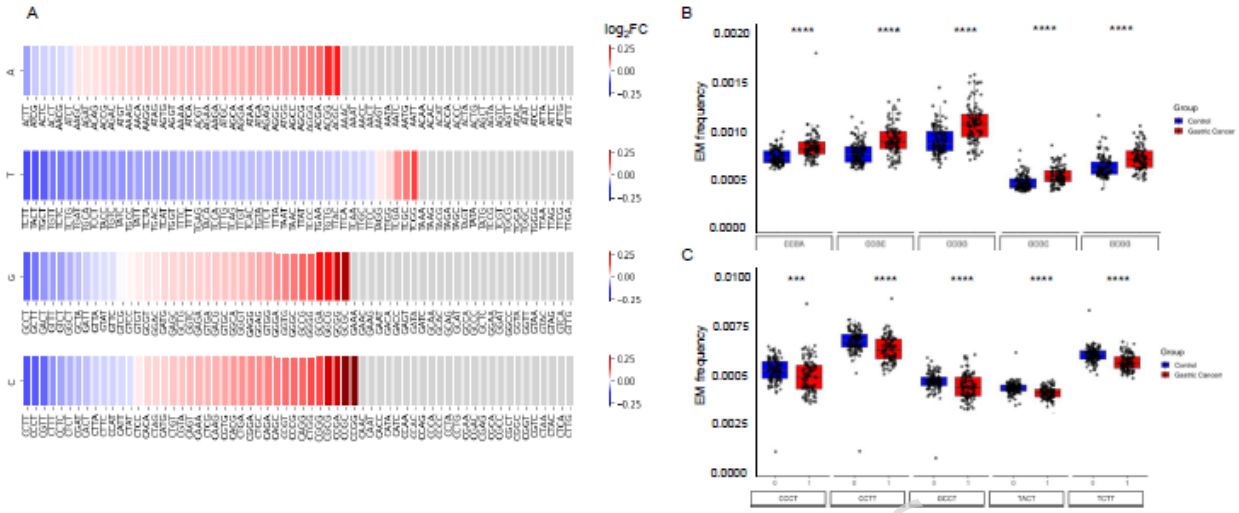

Figure 4

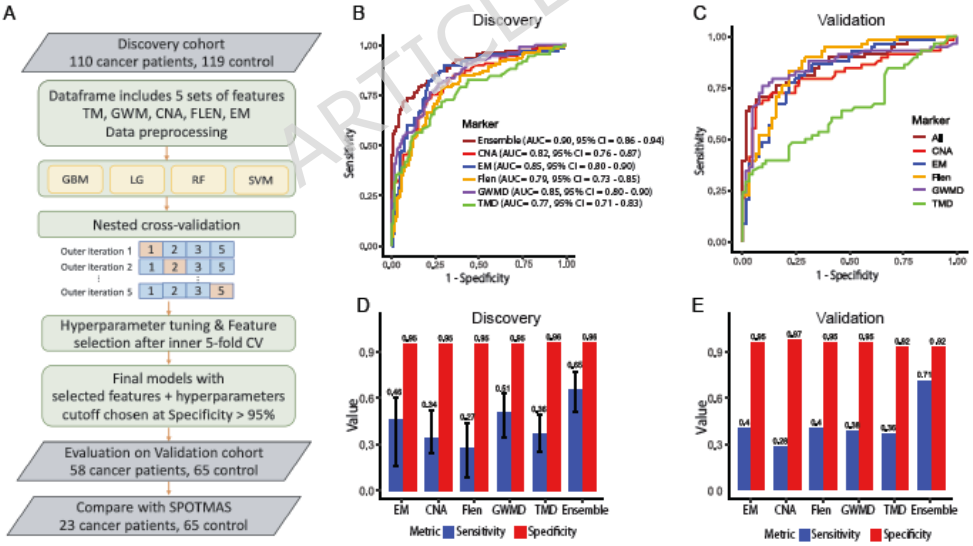

Figure 5

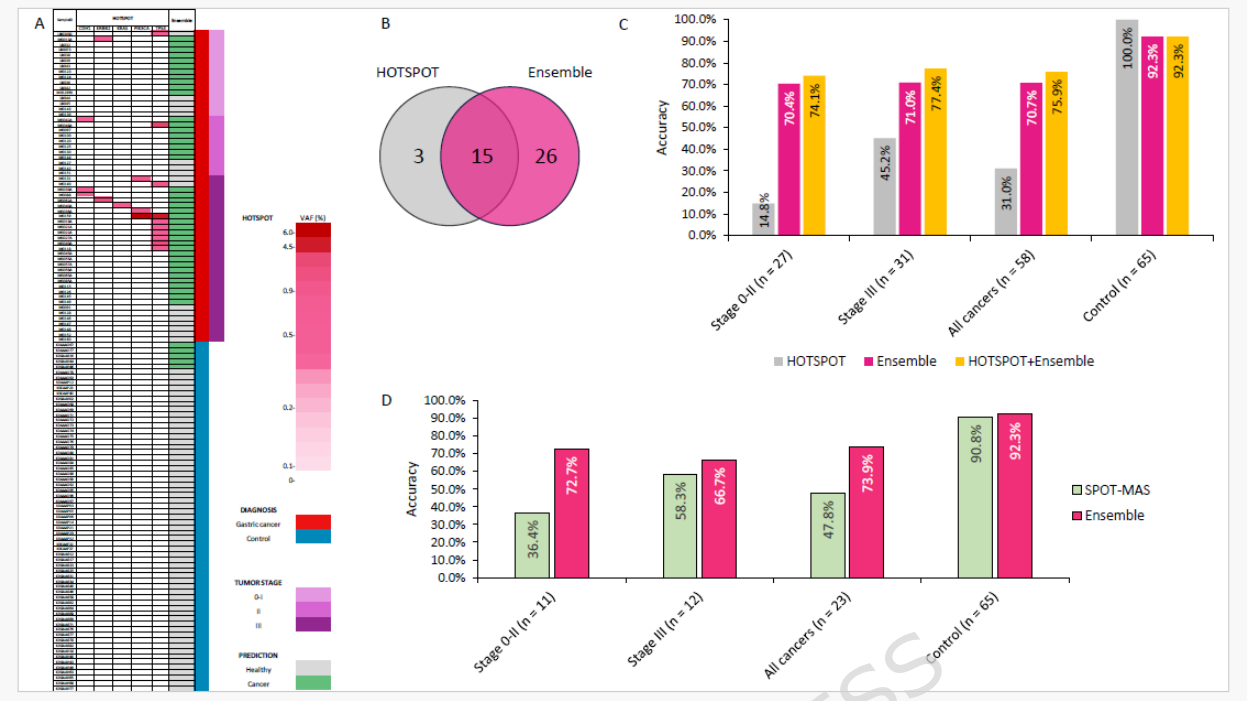

Figure 6

Table 1 Summary of clinical information of 204 gastric cancer patients and 284 non-cancer subjects (total N = 488)

Clinical features		Discovery cohort (N=229)					Validation cohort (N=123)					External validation cohort (N=136)					
		Gastric cancer (N = 110)		Healthy control (N = 119)		p-value (All cancer vs Healthy control)	Gastric cancer (N = 58)		Healthy control (N = 65)		p-value (All cancer vs Healthy control)	Gastric cancer (N = 36)		Healthy control (N = 71)		Gastritis (N = 29)	
		N	Percentage	N	Percentage		N	Percentage	N	Percentage		N	Percentage	N	Percentage	N	Percentage
Gender	Female	39	35.5%	50	42.0%	Chi-Square test p-value = 0.309	25	43.1%	32	49.2%	Chi-Square test p-value = 0.496	15	25.9%	40	61.5%	17	26.2%
	Male	71	64.5%	69	58.0%		33	56.9%	33	50.8%		21	36.2%	31	47.7%	12	18.5%
Age	Median	63		52		Mann-Whitney test p-value < 0.0001	59		45		Mann-Whitney test p-value < 0.0001	62		48		54	
	Min	36		28			32		25			35		29		26	
	Max	97		75			79		75			86		74		77	
Stage	0, IA, IB	39	35.5%				16	27.6%				14	38.9%				
	IIA, IIB	17	15.5%				11	19.0%				7	19.4%				
	IIIA	44	40.0%				31	53.4%				9	25.0%				
	Non-metastasis with unknown staging information	10	9.1%				0	0.0%				6	16.7%				