Breast Cancer Research

RESEARCH Open Access

Multidimensional cell-free DNA fragmentomics enables early detection of breast cancer

Lixian Yang[1,2†], Mengyang An[1†], Heng song[1], Xuan Zhang[2], Meiqi Wang[1], Liu Yang[1], Xinle Wang[1], Hua Yang[3], Xinyue Hong[4] and Zhenchuan Song[1,5*]

Abstract

Background Cell-free DNA (cfDNA) fragmentomics represents a transformative approach for early breast cancer detection, offering significant potential to improve patient survival through timely intervention. Despite this promise, existing cfDNA-based methods demonstrate inadequate sensitivity for clinical implementation, particularly in early-stage malignancies. There remains an urgent need to develop robust, cost-effective diagnostic strategies integrating cfDNA fragmentomic profiling with advanced machine learning algorithms.

Methods This research involved a total of 191 participants who did not have cancer and 204 participants diagnosed with breast cancer. The plasma cfDNA samples from the participants underwent profiling through whole-genome sequencing. A variety of cfDNA characteristics and machine learning models were assessed within the training cohort to attain the best model. The evaluation of model performance took place in a separate validation cohort.

Results An assembled ensemble model that combines three cfDNA characteristics with six machine learning algorithms, developed in the training cohort (cancer: 119; healthy: 112), outperformed all models created from individual feature-algorithm pairs. This composite model demonstrated enhanced sensitivities of 93.3% at a specificity of 94.6% for the training cohort (area under the curve [AUC], 0.983) and 96.5% at 93.7% specificity for the validation cohort (AUC, 0.989) (cancer: 85; healthy: 79). Additionally, our model exhibited sensitivity across various stages, distinct pathological types, and diverse molecular classifications.

Conclusion We have established a stacked ensemble model using cfDNA fragmentomics features and achieved superior sensitivity for detecting early-stage breast cancer, which could promote early diagnosis and benefit more patients.

Keywords Breast cancer, Early detection, Cell-free DNA, Whole-genome sequencing, Machine learning

†Lixian Yang and Mengyang An have contributed equally to this work.

*Correspondence:
Zhenchuan Song
songzhch@hebmu.edu.cn
[1]Breast Center, The Fourth Hospital of Hebei Medical University, 169 Tianshan Street, Shijiazhuang 050000, Hebei, People's Republic of China
[2]Department of Breast Surgery, Xingtai People's Hospital, No. 818 Xiangdu district, Xingtai 054000, Hebei, People's Republic of China
[3]Department of Medical oncology, Hebei University Affiliated Hospital, Shijiazhuang, People's Republic of China
[4]Geneseeq Research Institute, Nanjing Geneseeq Technology Inc, Nanjing 210032, People's Republic of China
[5]Key Laboratory for breast cancer molecular medicine of Hebei Province, Shijiazhuang 050035, Hebei, People's Republic of China

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Background

Breast cancer persists as a formidable global health challenge, accounting for approximately 2.3 million new cases and 650,000 deaths annually, with hormone receptor–positive/HER2-negative subtypes representing the predominant burden [1]. According to the American Cancer Society, an estimated 313,510 new cases and 42,780 deaths are expected in the United States alone in 2024 [2]. Current diagnostic strategies include clinical examination, mammography, tissue biopsy, and blood-based assays for tumor-associated antigens or proteins [3, 4]. Early and accurate detection remains critical for improving patient outcomes, highlighting the urgent need for reliable, breast cancer–specific biomarkers to identify aggressive disease phenotypes [5].

Despite notable advances in breast cancer diagnosis and treatment, conventional approaches remain constrained by several limitations, including invasiveness, high cost, and suboptimal sensitivity and specificity, which may limit their applicability across diverse patient populations [6]. Standard diagnostic modalities such as imaging and tissue biopsy are further challenged by risks of false-positive and false-negative results, radiation exposure, and their inability to comprehensively capture the genomic and molecular heterogeneity of tumors [7–9]. These limitations underscore the urgent need for novel, noninvasive strategies capable of enabling early detection, predicting prognosis, and monitoring therapeutic response.

In this context, recent studies have increasingly highlighted the importance of circulating molecular biomarkers detectable in peripheral blood. Among these, circulating tumor cells (CTCs), microRNAs, and cell-free DNA (cfDNA) have attracted considerable attention due to their noninvasive nature and reproducibility in liquid biopsy assays. The ease of obtaining these biomarkers, coupled with their potential for real-time monitoring of tumor dynamics, has facilitated their application in early detection and therapeutic management [10–13]. These biomarkers hold promise as both diagnostic and prognostic tools, providing insights into treatment response, drug resistance, and metastatic progression [14–16]. Moreover, accumulating evidence suggests that circulating biomarkers may play a pivotal role in advancing personalized medicine strategies for breast cancer patients [11, 14, 16–19].

Cell-free DNA (cfDNA) consists of short segments of nucleic acid found in plasma. At first, it was thought that these fragments originated from processes such as apoptosis or necrosis [20, 21]. Nonetheless, recent research suggests that cfDNA may be actively secreted by tumor cells [22–27] and might also be associated with circulating tumor cells (CTCs) present in the bloodstream [28]. A significant portion of studies on circulating cfDNA has focused on its potential as a liquid biopsy biomarker, either for assessing tumor burden through concentration analysis or for monitoring treatment response and prognosis by investigating genetic and epigenetic modifications. Techniques centered on methylation, such as whole-genome bisulfite sequencing and additional bisulfite modifications in next-generation sequencing (NGS) approaches, have emerged as promising methods for the early identification of cancer. Nonetheless, these techniques are associated with significant drawbacks, including procedural intricacies and prolonged timelines. As a result, there is an urgent need for other strategies that can mitigate these technical challenges and reduce the time required, while maintaining high sensitivity and specificity for the early detection of cancer.

The emerging field of fragmentomics, which investigates the patterns of circulating free DNA (cfDNA) fragmentation, is anticipated to address the existing gap caused by the insufficiency of comprehensive and detailed whole genome sequencing (WGS) data. Notably, the fragmentation patterns of cfDNA exhibit a heterogeneous distribution across the genomic landscape. These patterns are believed to be influenced by chromatin structure, offering essential insights into nucleosome arrangement and gene expression [13]. A notable breakthrough was achieved with the DNA Evaluation of Early Intercepted Fragments (DELFI), which assesses the coverage, size, and various summary statistics of fragments within a 5 Mb region [29]. The ratios of short (100–150 bp) and long (151–220 bp) cfDNA fragments were found to differ between healthy individuals and cancer patients, allowing for differentiation according to disease status. Furthermore, flagmenomics studies have integrated additional methods, including preferential terminal coordinates [30] and terminal motifs [31], to differentiate between cancerous and non-cancerous groups. Recently, analyses of transcription factors linked to cfDNA have been conducted to identify specific patterns unique to patients and tumors with significant accuracy [32].

Recognizing the valuable insights provided by circulating free DNA (cfDNA) and its capacity to distinguish between cancerous and non-cancerous states, we developed a model designed to encapsulate the intricate landscape of cfDNA. By utilizing five refined fragmentation attributes, we employed a robust integrated stacking machine learning methodology to construct the model. The primary objective of this study is to present a model that is both highly sensitive and economically viable for the identification of breast cancers (BCs), potentially leading to significant clinical benefits in real-world medical applications.

Materials and methods

Participant enrollment

For model construction and internal validation, we enrolled 395 participants in this study cohort, including 204 patients with breast cancer (BC) and 191 healthy participants, all sourced from the Fourth Hospital of Hebei Medical University. The present study was approved by the Medical Ethics Committee of the Fourth Hospital of Hebei Medical University (No. 2023KY032), and was carried out in accordance with the Declaration of Helsinki.

Whole genome sequencing for plasma samples

Ten milliliters of peripheral blood were drawn from participants into EDTA tubes. This was followed by centrifugation at 16,000 x g for a duration of 10 min within six hours to isolate the plasma. The resulting plasma samples were then stored at a temperature of -80 °C until they were transported on dry ice to the central laboratory at Nanjing Geneseq Technology Inc., a clinical testing facility located in China.

Upon arrival, the plasma samples were stored temporarily at -20 °C for roughly 1 to 3 days, depending on when the sequencer became available. Cell-free DNA (cfDNA) was isolated from the thawed plasma samples utilizing the QIAamp Circulating Nucleic Acid Kit (Qiagen). The cfDNA concentration was subsequently assessed with the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific), following the protocols outlined by the manufacturer. Following this, 5–10 ng of cfDNA was employed to construct whole genome sequencing (WGS) libraries, utilizing DNA end repair, A-tailing, and adapter ligation with the KAPA Hyper Prep Kit (KAPA Biosystems; according to the manufacturer's guidelines). The resulting libraries were quantified using KAPA SYBR FAST qPCR Master Mix (KAPA Biosystems) prior to being sequenced in paired-end mode on the T7 platform (MGI).

To maintain the integrity of the data, a thorough array of quality assurance procedures was implemented on the generated sequence information. The first procedure included the refinement of sequence reads with Trimmomatic (v0.36.6). Next, PCR duplicates were identified and removed using the Picard toolkit (http://broadinstitute.github.io/picard/). Once the trimming was completed, the sequences were aligned to the human reference genome (GRCh37/UCSC hg19) using the BWA aligner. QC metrics included Q30 scores, GC content, alignment rate to the reference genome, median insert size, and mean depth. Samples with sequencing QC failures (mean depth < 5×) were excluded. To ensure unbiased processing, the sample-operating team remained blinded to the case/control status throughout the workflow.

To address the variability in sequencing depths across samples, those with a depth of greater or equal to 5× were down-sampled to a uniform depth of 5× in order to optimize the model. The optimized model was subsequently validated using raw WGS data and additional down-sampling to depths of 4×, 3×, 2×, and 1× with the Picard toolkit, enabling a comprehensive evaluation of its performance.

CfDNA fragmentomics features extraction

We utilized three distinct fragmentomics feature derived from cfDNA shallow WGS data, including: *C*opy *N*umber *V*ariation (CNV), *F*ragment *S*ize *D*istribution (FSD), and *FRAG*ment based *M*ethylation *A*nalysis (FRAGMA). The total number of extracted features reached 3,432, including 2,475 for CNV, 936 for FSD, and 21 for FRAGMA.

The extraction of CNV features was conducted according to the methodology described by Wan et al. [33], utilizing the ichorCNA tool. The genome was segmented into 1 Mb non-overlapping bins (totaling 2475 segments) based on the GRCh37/UCSC hg19 reference genome. The analysis incorporated multiple normalization steps: (1) GC content correction using precomputed GC wig files to address sequencing biases, (2) mappability adjustment with genome-specific map wig files to account for alignment artifacts in repetitive regions, and [20] systematic noise reduction through a Panel of Normals (PoN) constructed from healthy donor cfDNA samples to eliminate technical artifacts from library preparation and sequencing platforms. Log2 ratios were computed after these corrections, with autosomes serving as the diploid baseline (log2 ratio = 0). Regarding the details, we first use the readCounter tool from the HMMcopy suite (v0.1.1) to calculate the read counts in different regions of the genome. Then, we use runIchorCNA.R to correct for GC content and mappability, thereby obtaining the final CNV results.

For the **F**ragment **S**ize **D**istribution (FSD), we adapted the methodology from Cristiano et al. [29]. The FSD profile was generated by analyzing cfDNA fragments within the 110–220 bp range at 5-bp intervals (24 size bins per chromosome arm), covering 39 chromosome arms (totaling 936 features) while excluding the short arms of five acrocentric chromosomes (13p, 14p, 15p, 21p, 22p) due to limited sequencing coverage. Raw fragment counts were first corrected for GC-content bias using LOESS regression, then standardized to z-scores $[(x-\mu)/\sigma]$ within each size bin across all samples, where x represents the GC-corrected fragment count, μ the mean, and σ the standard deviation. This dual normalization approach effectively removed technical artifacts while enhancing biologically relevant fragmentation patterns for improved discrimination between cancerous and noncancerous cases.

The FRAGmentomics-based Methylation Analysis (FRAGMA) was introduced by Qing Z. et al. utilized

cfDNA cleavage patterns around cytosine-phosphate-guanine (CpG) sites [34]. This methodology argued that an elevated likelihood of cfDNA cleavage adjacent to cytosine is indicative of methylation at CpG sites, whereas such a pattern is generally absent at unmethylated sites. Alu regions were specifically targeted where methylation levels are notably higher compared to the whole genome and CpG islands. Within these Alu regions, the fragment ratios of 5′ end motifs of CGN and NCG (N represents any nucleotide) resulted in eight fragment ratios. Additionally, the ratio of CGN to NCG motifs were computed. Notably, an enhanced rate of cleavage at the two cytosine positions was observed in CGCG sequence. Expanding this analysis, every pair of CpG nucleotides within the CGCG sequence was examined. The resulting fragment ratios were derived from 5' end motifs of CGC, NCG, CGN, and GCG identified, resulting in ten fragment ratios for the CGCG sequence. Finally, the motif ratios of CGC/NCG and CGN/CGC were calculated.

Machine learning model construction

The dataset used for training included a total of 119 patients with breast cancer and 112 healthy individuals, while the independent validation dataset was kept separate from the training process and was utilized solely for assessing the final model.

To construct predictive models, the input data for the model consisted of 3432 distinct feature columns (CNV: 2,475, FSD: 936, FBM:21) without further feature selection, ensuring that all available cfDNA characteristics contributed to both the cancer classifier model and the tumor of origin classifier model. During the modelling process, each fragmentomics feature type was used to build individual base models by applying six machine learning algorithms – Generalized Linear Model (GLM), Extremely Randomized Trees (XRT), Deep Regression Forests (DRF), Deep Learning [35], Gradient Boosting Model (GBM), and eXtreme Gradient Boosting (XGBoost). These algorithms helped generate scores ranging from 0 to 1 as the cancer probability indicator. Higher scores implied greater probability of cancer. The final stacked ensemble model was created by averaging the cancer scores from all algorithms. The model training and stacking process were completed by H2O AutoML framework (v3.36.0.3) [36]. This approach employed a random grid search to automatically optimize comprehensive hyperparameters across models, including booster type, maximum tree depth, column sampling, minimum rows per leaf node, number of trees, and regularization parameters (reg_alpha, reg_lambda). Importantly, the model was trained exclusively on the training dataset, ensuring the integrity of the validation set until the final performance evaluation. After training the base models, H2O AutoML ranked all models on a Leaderboard based on their area under the curve (AUC)metrics. Notably, 5-fold cross-validation was applied to both the based models and the stacked ensemble model. The training cohort was divided into five subsets, with four used for training and one for validation in each iteration. The model robustness was then ensured by repeating this process across all subsets. Additionally, base model training limits and ensemble techniques were incorporated in our study to enhance generalizability. Following performance evaluation of all models, we randomly combined the combined features with the top AUC from 6 algorithms to obtain the best-performing stacked model. Then, this optimal model was used to predict each sample in the validation cohort. The cut-offs were determined by training cohort at 95% specificity for the cancer classifier.

All statistical analysis

The ROC curves were created using the pROC package (version 1.18.5). We calculated the sensitivity [TP/(TP + FN)], specificity [TN/(TN + FP)], positive predictive value (PPV) [TP/(TP + FP)], negative predictive value (NPV) [TN/(TN + FN)], and accuracy [(TP + TN)/(TP + FP + TN + FN)], as well as the corresponding 95% confidence intervals, based on true-positive (TP), true-negative [9], false-positive (FP), and false-negative [33] values using the epiR package (version 2.0.75). All statistical analyses were performed in R (version 4.4.1).

Results

Participants' disposition and characteristics

The dataset utilized for training the model consisted solely of 231 individuals (119 BC patients and 112 healthy controls). For the validation cohort, 85 BC patients and 79 healthy individuals were included. Samples for this validation cohort were systematically gathered from the same center during the period of April 2023 to July 2023 (Fig. 1A).

Table 1 provides a summary of the demographics and clinical features for all participants in both the training and validation cohorts. In the training cohort, the average age of BC patients was 54 years, while in the validation cohort, the average age was 56 years. For healthy participants, the mean ages were 54 years in the training cohort and 55 years in the validation cohort. The cancer stage distribution remained similar between the cohorts, with most participants classified as stage II disease (training cohort: 52.4%; validation cohort: 50.6%). The participant group consisted entirely of females.

Overall, the demographic characteristics showed similarities between the training and validation cohorts, ensuring reliable performance estimates.

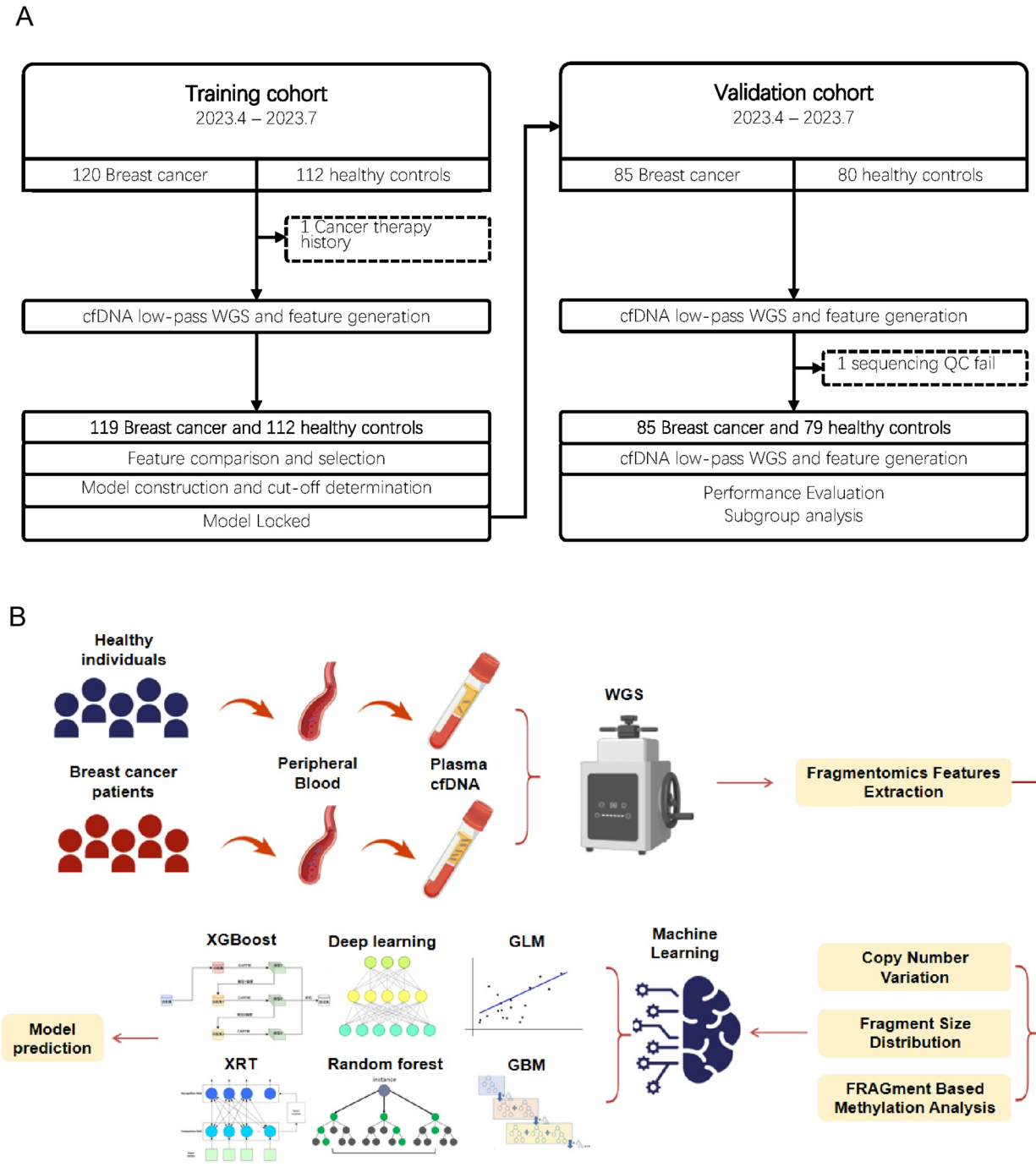


Fig. 1 Dataset and model construction workflow **A** From 2023.04 to 2023.07, 119 breast cancer patients and 112 healthy participants were enrolled for the training dataset. To evaluated model performance, from 2023.04 to 2023.07, 85 breast cancer patients and 79 healthy participants were enrolled. **B** Schematic diagram of the stacked ensemble model construction and cancer probability score determination. Plasma samples went through the low-pass whole-genome sequencing (WGS) process, and the machine learning prediction model was constructed based on three cell-free DNA (cfDNA) fragmentomics features: Copy Number Variation (CNV), Fragment Size Distribution (FSD) and Fragment Methylation Analysis (FragMA). The genome-wide feature profiles were then applied to the six machine learning algorithms, with the resultant matrix processed by a second-layer GLM algorithm to form the stacked ensemble model and calculate the participant's cancer probability score

Table 1 Baseline characteristics of the sample population

	Training dataset		Validation dataset	
	Breast cancer	**Healthy**	**Breast cancer**	**Healthy**
	***N*=119**	***N*=112**	***N*=85**	***N*=79**
Age	56.2 [30, 79]	54.0 [45, 74]	52.7 [28, 74]	53.2 [45, 70]
Sex, percentage				
Female	119 (100%)	112 (100%)	85 (100%)	79 (100%)
TNM Stage, percentage				
Stage I	50 (42.0%)	N/A	23 (27.1%)	N/A
Stage II	59 (49.6%)	N/A	52 (61.2%)	N/A
Stage III	10 (8.4%)	N/A	10 (11.8%)	N/A
Histology type, percentage				
Ductal	99 (83.2%)	N/A	69 (81.2%)	N/A
Papillary	9 (7.6%)	N/A	4 (4.7%)	N/A
Lobular	4 (3.4%)	N/A	8 (9.4%)	N/A
other	7 (5.9%)	N/A	4 (4.7%)	N/A
Differentiation, percentage				
Moderate	83 (69.7%)	N/A	59 (69.4%)	N/A
Poor	27 (22.7%)	N/A	18 (21.1%)	N/A
Unknown	9 (7.6%)	N/A	8 (9.4%)	N/A
Molecular subtype, percentage				
HR- & HER+	77 (64.7%)	N/A	49 (57.6%)	N/A
HER+ (HR + or HR-)	30 (15.2%)	N/A	21 (24.7%)	N/A
TNBC	10 (8.4%)	N/A	12 (14.1%)	N/A
Unknown	2 (1.7%)	N/A	3 (3.5%)	N/A

Feature assessment and ensemble stacked model evaluation

Feature assessment and ensemble stacked model evaluationModel selection was conducted by assessing the AUC values obtained from various combinations of the cfDNA features and machine learning algorithms across the validation cohorts. We evaluated the performance of FRAGMA, FSD, and CNV features within all foundational models (including GLM, GBM, deep learning, random forest, and XGBoost) in terms of their AUC metrics. The analysis of different foundational models demonstrated that the algorithm-stacked model outperformed others, as the integration of FRAGMA, FSD, and CNV features led to higher AUC values within the stacked model compared to those from individual algorithm models. Consequently, we developed a stacked ensemble model that incorporates the three plasma cfDNA fragmentomic features (FRAGMA, FSD, and CNV) alongside six machine learning algorithms (XRT, GLM, GBM, random forest, deep learning, and XGBoost) (see Fig. 1B). The stacked ensemble model achieved an enhanced predictive capability, resulting in an AUC of 0.983, which surpassed the performance of stacked models that utilized individual fragmentomic features (Fig. 2A). Furthermore, the AUC of the stacked ensemble model exceeded that of a generalized linear model (GLM) based on fragmentomic features. Consequently, we selected the stacked ensemble model for subsequent evaluations. Our analysis revealed that the predicted cancer scores for cancer patients were significantly greater than those of healthy individuals within the training cohort (Fig. 2B). In the training set, the model demonstrated sensitivity and specificity rates of 93.3% and 94.6%, respectively, in differentiating breast cancer from healthy subjects (Fig. 2C). Figure 2D illustrates the prediction scores of the model for both breast cancer and healthy samples, additionally highlighting false detections that occurred under the 95% specificity threshold (0.504).

We evaluated the effectiveness of the stacked ensemble model using the validation dataset. The area under the curve (AUC) of the model remains notably high, measuring 0.989 in the validation cohort (Fig. 3A). The median scores for samples collected from breast cancer patients are considerably greater compared to those from healthy individuals, yielding a Wilcoxon test p-value of under 0.01 (Fig. 3B).

Performance of the predictive model in different cancer sample subgroups

The performance of the model was additionally assessed across various breast cancer subgroups utilizing the validation datasets. Notably, the subgroups exhibited no statistically significant differences among the various categories (Fisher's exact test, P values exceeding 0.05). As illustrated in Fig. 3C-F, the sensitivity of assay detection remains consistently elevated among different patient subgroups with breast cancer. Specifically, our model demonstrated sensitivity across various stages, pathologic types, levels of differentiation, and molecular classifications. It was important to note that moderate differentiation corresponded to histological grades II and III, whereas low differentiation was identified as histological grade I.

Discussion

Emerging as a non-invasive supplement or substitute for traditional tumor biopsies, liquid biopsies are gaining prominence [37–41]. In this research, we present a stacked ensemble model constructed from three cfDNA characteristics (FSD, CNV, FRAGMA) utilizing six distinct algorithms (GLM, GBM, DL, XRF, DRF, and XGBoost). This model serves as an ultra-sensitive and cost-effective non-invasive method for breast cancer detection. Furthermore, our findings indicate potential advantages in stage shift at diagnosis when applied in real-world scenarios within Chinese demographics. Our objective was to improve the early identification of breast cancer by developing a stacked ensemble machine learning model that integrates various cfDNA fragmentomic

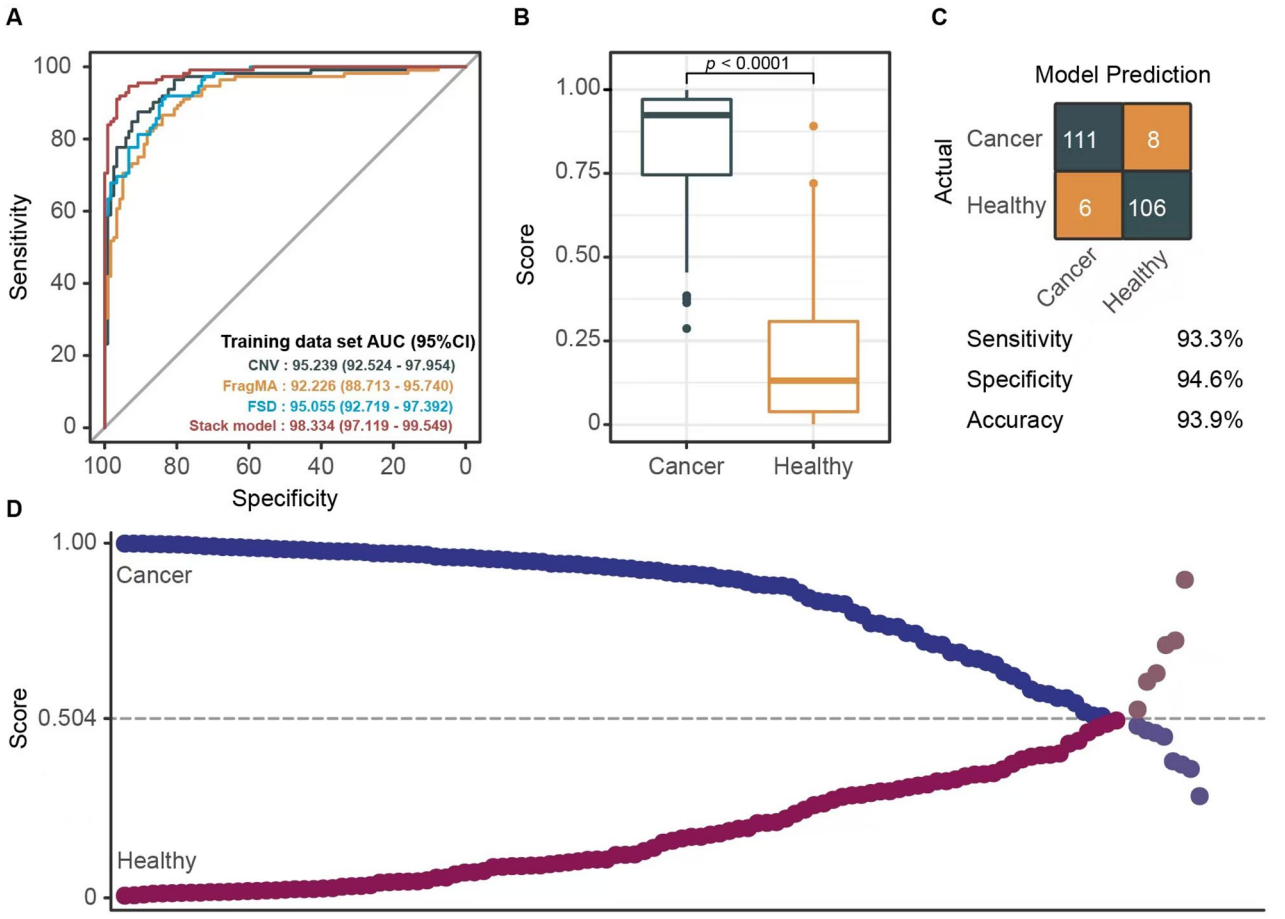


Fig. 2 Model construction in training dataset. **A** Receiver operating characteristic (ROC) curves evaluating the model's predictive performance of base model (CNV, FSD, FragMA) and stack model. **B** The median score of samples from breast cancer patients is significantly higher than that of healthy individuals, with a Wilcoxon test p-value of less than 0.01. **C** The confusion matrix, sensitivity, specificity, and accuracy of the ensemble stack prediction model in the training dataset. **D** Scatterplot showing the model prediction scores for breast cancer and healthy samples, as well as labelling the false detections under the 95% specificity cut-off (0.504).

features, thereby achieving high sensitivity in differentiating individuals with early-stage breast cancer from those without the disease.

Our multidimensional assay leverages the advantages of cfDNA fragmentomic characteristics within a stacked ensemble machine learning framework. The cleavage and fragmentation of cfDNA occur in a non-random manner, with certain genomic regions exhibiting preferred cleavage patterns, known as preferred end sites. These patterns are associated with factors such as tissue origin and disease states, which are influenced by chromatin accessibility, nuclease activities, and other elements [42].Given that preferred end sites linked with tumors are more widespread and easily detected than mutations, fragmentomic features have emerged as a significant category of ctDNA signatures [42]. The exceptional performance of our model underscores the potential of cfDNA fragmentomics in cancer detection. As previously mentioned, patterns of cfDNA methylation and fragmentation have been employed to identify breast cancer patients; however, their sensitivity for clinical application remains suboptimal [14, 42–45]. We have conducted a comprehensive assessment of the current characteristics of circulating free DNA (cfDNA) and integrated the most effective attributes into our predictive model. Ensemble machine learning models exhibit significant advantages in terms of accuracy, reproducibility, and interpretability for bioinformatics applications. Additionally, the stacking approach effectively consolidates the predictions produced by individual base models [44].

Focusing exclusively on the whole genome sequencing (WGS) of plasma cell-free DNA (cfDNA), our research presents an exemplary performance model for cancer identification. It attains an area under the curve (AUC) of 0.9833 in the independent validation cohort and reaches an AUC of 0.9895 in prospective evaluations. This remarkable predictive capability is due to a comprehensive ensemble stacked model utilizing machine learning

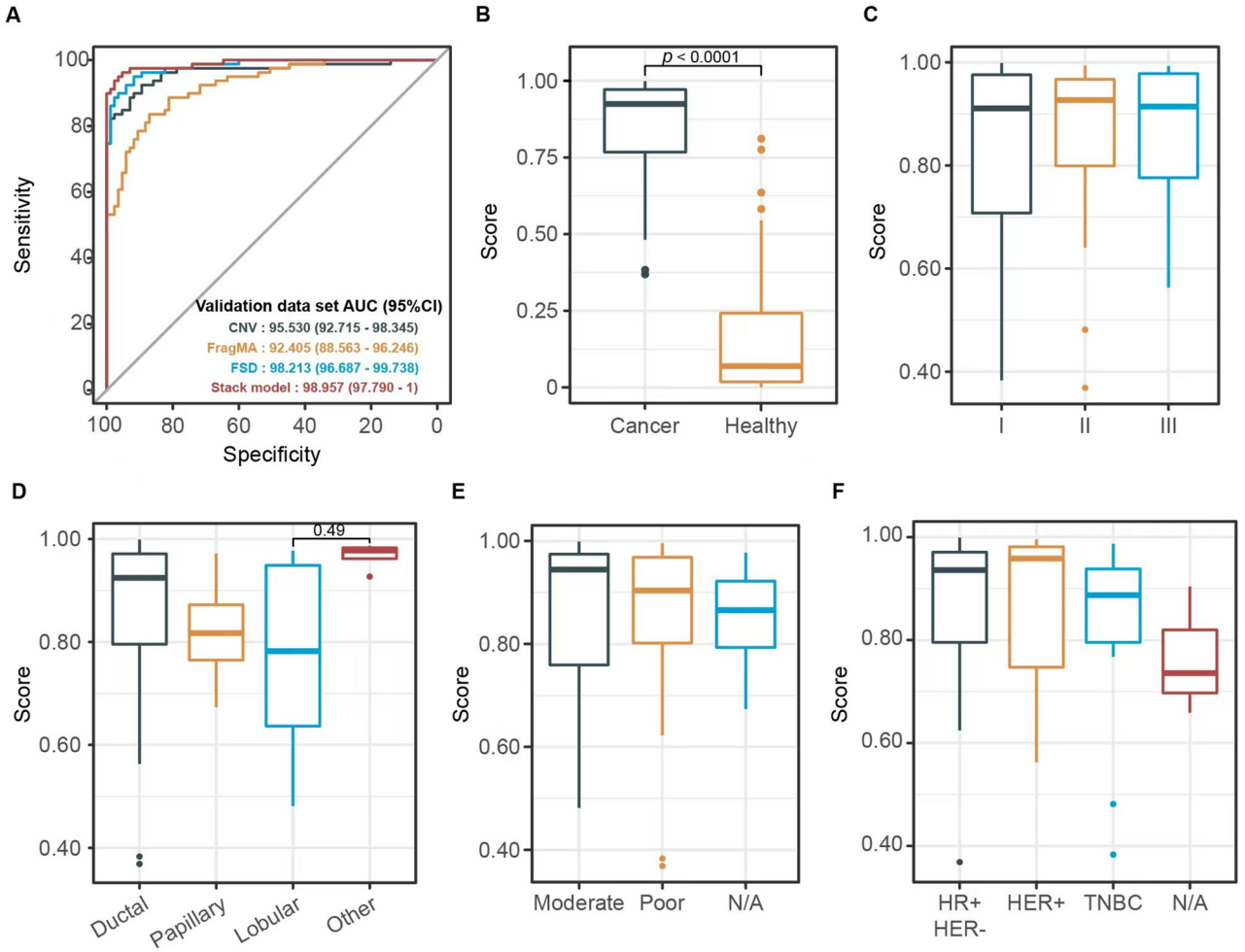


Fig. 3 Model performance evaluation and subgroup analysis in validation dataset. **A** Receiver operating characteristic (ROC) curves evaluating the model's predictive performance of base model (CNV, FSD, FragMA) and stack model. **B** The median score of samples from breast cancer patients is significantly higher than that of healthy individuals, with a Wilcoxon test p-value of less than 0.01 **C**-**F** Prediction score differences of breast cancer samples, stratified by histological subtype (**D**), differentiation level (**E**) and molecular classification (**F**)

techniques and relying on three cfDNA characteristics. Importantly, no individual-feature base model was able to achieve comparable outcomes. The ensemble stacked model's error correction features improve overall predictive accuracy (for instance, accurately classifying non-cancer individuals as non-cancer and cancer individuals as cancer). This advancement leads to heightened sensitivity and specificity, yielding a superior AUC in comparison to the base models. Furthermore, it is important to note that this model exhibits strong performance across various subgroups, including tumor stage, histological grade, and molecular classification.

This study highlights several limitations. While we have demonstrated the application of cfDNA fragmentomics for enhanced sensitivity in detecting breast cancer, the underlying mechanisms in this field remain poorly understood. We are currently conducting an extensive inquiry aimed at identifying factors and revealing new characteristics that could refine our model. Furthermore, the sample size is somewhat limited, which may lead to an imprecise evaluation of sensitivity within smaller subgroups, including tumors with low or high differentiation. Moreover, we were unable to include a fully external cohort from independent institutions in the current analysis. We recognize that such external validation is essential to firmly establish clinical utility and minimize potential biases. Moving forward, we are actively planning a prospective, multi-center study to evaluate the assay's performance in a real-world screening context.

Conclusions

In conclusion, we created a stacked ensemble model that integrates three fragmentomic characteristics obtained from plasma cfDNA whole genome sequencing (WGS) data. Our model showed remarkable sensitivity in distinguishing patients with early-stage breast cancer from

those without the condition. In addition to improving the ability to detect very early and small tumors in breast cancer patients, our comprehensive model provides an accurate and economical approach for promoting early breast cancer detection and improving patient outcomes.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13058-025-02190-8.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Acknowledgements
Not applicable.

Author contributions
Z.S. conceived the study. LX.Y. and M.A. designed the methodological framework. H.S., X.Z. and X.H. performed the data analysis. M.W., L.Y., H.Y., and X.W. wrote the manuscript together. All the authors read and approved the final manuscript.

Funding
This work was supported by the following funding: Supported by Hebei Natural Science Foundation (No. H2020206365), S&T Program of Hebei (No. 236Z7719G) and Innovative Drugs Post-market Clinical Research Special Program (No. WKZX2024CX103205).

Data availability
The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate
The present study was approved by the Medical Ethics Committee of the Fourth Hospital of Hebei Medical University (No. 2023KY032), and was carried out in accordance with the Declaration of Helsinki.

Consent for publication
All authors have consented to the publication of this article.

Competing interests
The authors declare no competing interests.

Received: 8 January 2025 / Accepted: 29 November 2025
Published online: 09 December 2025

References

1. Bray F, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229–63.
2. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. CA Cancer J Clin. 2024;74(1):12–49.
3. Weigelt B, et al. Refinement of breast cancer classification by molecular characterization of histological special types. J Pathol. 2008;216(2):141–50.
4. Shaw JA, Stebbing J. Circulating free DNA in the management of breast cancer. Ann Transl Med. 2014;2(1):3.
5. Saadatmand S, et al. Influence of tumour stage at breast cancer detection on survival in modern times: population based study in 173,797 patients. BMJ. 2015;351:h4901.
6. Alimirzaie S, Bagherzadeh M, Akbari MR. Liquid biopsy in breast cancer: A comprehensive review. Clin Genet. 2019;95(6):643–60.
7. Palmirotta R, et al. Liquid biopsy of cancer: a multimodal diagnostic tool in clinical oncology. Ther Adv Med Oncol. 2018;10:1758835918794630.
8. Marinovich ML, et al. Agreement between digital breast tomosynthesis and pathologic tumour size for staging breast cancer, and comparison with standard mammography. Breast. 2019;43:59–66.
9. Zubor P et al. Why the gold standard approach by mammography demands extension by multiomics? Application of liquid biopsy MiRNA profiles to breast cancer disease management. Int J Mol Sci, 2019. 20(12).
10. Schwarzenbach H, Hoon DS, Pantel K. Cell-free nucleic acids as biomarkers in cancer patients. Nat Rev Cancer. 2011;11(6):426–37.
11. Umetani N, et al. Prediction of breast tumor progression by integrity of free Circulating DNA in serum. J Clin Oncol. 2006;24(26):4270–6.
12. Giuliano M, et al. Circulating tumor cells as early predictors of metastatic spread in breast cancer patients with limited metastatic dissemination. Breast Cancer Res. 2014;16(5):440.
13. Snyder MW, et al. Cell-free DNA comprises an in vivo nucleosome footprint that informs its Tissues-Of-Origin. Cell. 2016;164(1–2):57–68.
14. Madhavan D, et al. Plasma DNA integrity as a biomarker for primary and metastatic breast cancer and potential marker for early diagnosis. Breast Cancer Res Treat. 2014;146(1):163–74.
15. Dawson SJ, et al. Analysis of Circulating tumor DNA to monitor metastatic breast cancer. N Engl J Med. 2013;368(13):1199–209.
16. Ma F, et al. CtDNA dynamics: a novel indicator to track resistance in metastatic breast cancer treated with anti-HER2 therapy. Oncotarget. 2016;7(40):66020–31.
17. Cheng J, et al. Circulating free DNA integrity and concentration as independent prognostic markers in metastatic breast cancer. Breast Cancer Res Treat. 2018;169(1):69–82.
18. Tang Z, et al. Diagnostic value of serum concentration and integrity of Circulating Cell-Free DNA in breast cancer: A comparative study with CEA and CA15-3. Lab Med. 2018;49(4):323–8.
19. Lin SY, Orozco JIJ, Hoon DSB. Detection of minimal residual disease and its clinical applications in melanoma and breast cancer patients. Adv Exp Med Biol. 2018;1100:83–95.
20. Choi JJ, Reich CF 3rd, and, Pisetsky DS. The role of macrophages in the in vitro generation of extracellular DNA from apoptotic and necrotic cells. Immunology. 2005;115(1):55–62.
21. Francis G, Stein S. Circulating Cell-Free tumour DNA in the management of cancer. Int J Mol Sci. 2015;16(6):14122–42.
22. Gahan PB, Swaminathan R. Circulating nucleic acids in plasma and serum. Recent developments. Ann N Y Acad Sci. 2008;1137:1–6.
23. Stroun M, et al. About the possible origin and mechanism of Circulating DNA apoptosis and active DNA release. Clin Chim Acta. 2001;313(1–2):139–42.
24. Gahan PB, Stroun M. The virtosome-a novel cytosolic informative entity and intercellular messenger. Cell Biochem Funct. 2010;28(7):529–38.
25. Bronkhorst AJ, et al. Characterization of the cell-free DNA released by cultured cancer cells. Biochim Biophys Acta. 2016;1863(1):157–65.
26. Stroun M, et al. The origin and mechanism of Circulating DNA. Ann N Y Acad Sci. 2000;906:161–8.
27. Wang W, et al. Characterization of the release and biological significance of cell-free DNA from breast cancer cell lines. Oncotarget. 2017;8(26):43180–91.
28. Schwarzenbach H, et al. Cell-free tumor DNA in blood plasma as a marker for Circulating tumor cells in prostate cancer. Clin Cancer Res. 2009;15(3):1032–8.
29. Cristiano S, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature. 2019;570(7761):385–9.
30. Jiang P, et al. Preferred end coordinates and somatic variants as signatures of Circulating tumor DNA associated with hepatocellular carcinoma. Proc Natl Acad Sci U S A. 2018;115(46):E10925–33.
31. Jiang P, et al. Plasma DNA End-Motif profiling as a fragmentomic marker in Cancer, Pregnancy, and transplantation. Cancer Discov. 2020;10(5):664–73.
32. Ulz P, et al. Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. Nat Commun. 2019;10(1):4666.
33. Wan N, et al. Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA. BMC Cancer. 2019;19(1):832.
34. Zhou Q, et al. Epigenetic analysis of cell-free DNA by fragmentomic profiling. Proc Natl Acad Sci U S A. 2022;119(44):e2209852119.
35. Becker PS, et al. Selection and expansion of natural killer cells for NK cell-based immunotherapy. Cancer Immunol Immunother. 2016;65(4):477–84.

36. LeDell E, Poirier S. *H2o automl: Scalable automatic machine learning*. in *Proceedings of the AutoML Workshop at ICML*. 2020.
37. Rothwell DG, et al. Utility of ctdna to support patient selection for early phase clinical trials: the TARGET study. Nat Med. 2019;25(5):738–43.
38. Ye Q, et al. Liquid biopsy in hepatocellular carcinoma: Circulating tumor cells and Circulating tumor DNA. Mol Cancer. 2019;18(1):114.
39. Lim SY, et al. Liquid biomarkers in melanoma: detection and discovery. Mol Cancer. 2018;17(1):8.
40. Xu RH, et al. Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. Nat Mater. 2017;16(11):1155–61.
41. Luo H et al. Circulating tumor DNA methylation profiles enable early diagnosis, prognosis prediction, and screening for colorectal cancer. Sci Transl Med, 2020. 12(524).
42. Adusei E et al. Reduced serum circulation of Cell-Free DNA following chemotherapy in breast cancer patients. Med Sci (Basel), 2021. 9(2).
43. Kamel AM, et al. Plasma DNA integrity index as a potential molecular diagnostic marker for breast cancer. Tumour Biol. 2016;37(6):7565–72.
44. Pham TMQ, 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;13:1127086.
45. Lamminaho M et al. *High Cell-Free DNA Integrity Is Associated with Poor Breast Cancer Survival.* Cancers (Basel), 2021. 13(18).

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.