ISSN: 2583-9888
DOI: doi.org/10.51219/JAIMLD/talha-waseem/677

URF PUBLISHERS
connect with research world

Journal of Artificial Intelligence, Machine Learning and Data Science

https://urfpublishers.com/journal/artificial-intelligence

Vol: 9 & Iss: 2

Research Article

Leveraging Artificial Intelligence for Multi-Cancer Early Detection Using Bloodstream Biomarker Analysis

Talha Waseem[1]*, Nisa Irshad[2], Irshad Mohammad[3], Dr. Mian Khurram Hanif[4], Tabeer Ali[5], Ankith Madireddi[6], Dr. Maqbool Khan[7] and Habiba Nasim[8]

[1]AI Research Scholar Program, Research, New York Institute of Technology College of Osteopathic Medicine, Long Island, New York

[2]AI Research Scholar Program, Research, Half Hollow Hills School District, Long Island, New York

[3]Assistant Director AI Research Scholar Program, Research, Genai-training.com LLC, NYC, New York

[4]Director Research & Lead AI Scholar Program, Research, Genai-training.com LLC, NYC, New York

[5]AI Research Scholar Program, Research, William L. Dickinson High School Jersey City, New Jersey

[6]AI Research Scholar Program, Research Washtenaw Technical Middle School, Ann Arbor, Michigan

[7]AI Research Scholar Program, Research, Assistant Professor, School of Computing Sciences, Pak-Austria Fach Hochschule: Institute of Applied Sciences and Technology, Haripur, KPK, Pakistan

[8]Health Informatics Specialist, AI Research Scholar Program, Research, Genai-training.com LLC, NYC, New York

Citation: Waseem T, Irshad N, Mohammad I, Hanif MK, Ali T, et al. Leveraging Artificial Intelligence for Multi-Cancer Early Detection Using Bloodstream Biomarker Analysis. *J Artif Intell Mach Learn & Data Sci 2026* 9(2), 3393-3400. DOI: doi.org/10.51219/JAIMLD/talha-waseem/677

Received: 08 March, 2026; **Accepted:** 19 April, 2026; **Published:** 27 April, 2026

***Corresponding author:** Talha Waseem, AI Research Scholar Program, Research, New York Institute of Technology College of Osteopathic Medicine, Long Island, New York

ABSTRACT

Early cancer detection is the cornerstone of management in neoplastic diseases. Traditional diagnostic techniques frequently miss cancers in their initial stages; therefore, early detection continues to be a major challenge in clinical oncology. Through empirical examination of bloodstream biomarkers-such as cell-free and circulating tumor DNA (cf/ctDNA), circulating tumor cells (CTCs), extracellular vesicles and RNA/DNA methylation profiles—early detection is now possible well before clinical expression or conventional screening via invasive biopsies, imaging, serology and antigen testing. This research paper explores the potential of artificial intelligence (AI) and machine learning (ML) to improve multi-cancer early detection (MCED). Three hypotheses-Diagnostic Accuracy, Early Signals and Pattern Recognition & Multi-Cancer Classification-were established and tested using methodologies such as Natural Language Processing (NLP), Deep Learning, Convolutional Neural Networks (CNN) and Large Language Model (LLM)-based classifiers. These AI feature-extraction approaches were utilized to identify hidden biomarker patterns and assess their potential correlation with early-stage disease and they underwent thorough validation. The results suggest that the framework improves sensitivity and specificity for detecting a variety of cancer types using minimally invasive blood samples. These findings provide strong evidence that AI-driven biomarker analysis can be incorporated into clinical practice, offering a scalable and empirically supported approach for prompt intervention and improved patient outcomes.

To further enhance the precision and therapeutic utility of multi-cancer early detection (MCED) systems, future research should focus on expanding the range of biomarkers, integrating multi-omics datasets and developing more generalized AI models.

Keywords: Artificial Intelligence & Machine Learning, Biomarkers, cf/ctDNA, Multi-Cancer Early Detection (MCED)

1. Introduction

The projected number of new cancer cases in the United States in 2025 is 2,041,910, with an estimated 618,120 cancer related deaths. The mortality rate from cancer has declined since 2022 but significant disparities persist among certain population groups. Native American and African American populations are approximately twice as likely to develop malignancies such as kidney, liver and stomach cancer in comparison to Caucasians. The incidence rate of cancer in women has also seen a rise[1]. Early detection is critical to further reducing these numbers. According to recent estimates, a 15% decline in cancer related deaths in the next 5 years can be accomplished through early disease detection[2]. At a later stage, these cancers are more aggressive and chemotherapy may be more toxic[3]. Furthermore, the increased costs associated with later detection and treatment is more likely to affect marginalized groups. The cost of treating advanced stage cancers (Stage III/IV) is substantially higher than that of early-stage cancers (Stage 1/II)[4]. This cost differential is also associated with changes in hospital stay and outpatient visits as a result for those with higher stage cancers[5].

Current strategies for cancer detection, although effective, lack patient adherence and are inaccessible to many. One of the main strategies is tissue biopsy. While this approach has been the diagnostic mainstay for decades, it has several limitations. It does not consistently detect cancers at sufficiently early stages. It can also be invasive or difficult to obtain based on anatomical position. Cancer screening tools such as mammograms, colonoscopies and low dose CTs are also some of the gold standards lacking some capabilities. Colonoscopies, although very accurate, are invasive and have relatively lower patient adherence rates[6]. A study done assessing social risk and nonadherence for recommended cancer screening, found that nonadherence was higher for colorectal and lung cancer in those that had public health insurance as compared to their counterparts with private coverage. Overall, they also found a higher prevalence of nonadherence for subjects that had social risks[7]. Nonadherence may not be solely attributable to social factors but may also be due to personal choice to avoid invasive testing. To combat this issue of noncompliance and late detection new strategies have been researched heavily.

This emerging strategy is based upon liquid biopsies which can be found in bodily fluids such as urine, saliva and blood[8]. These biopsies can provide an indication of possible early cancer development. The biomarkers that have been studied the most extensively are circulating tumor DNA (cfDNA/ctDNA), circulating tumor cells (CTCs), microRNA (miRNA) and exosomes. Biomarkers shed their analytes in bodily fluids early in the course of cancer development. Levels of these biomarkers have been found to be elevated in multiple cancer types such as prostate, colorectal, stomach, lung etc. The benefit of this strategy is that it can have the capacity to detect possible cancers before any physical masses or symptoms arise[9]. Detection can be achieved across a broad range of cancer types and is relatively efficient. Multicancer early detection tests have been developed as a cost-effective approach[10]. The application of Artificial Intelligence and Machine Learning will only speed up the process more by providing efficient data analysis. It can detect whether cancer is present or not in up to a few seconds or minutes.

Machine Learning models can learn subtle features long before cancer is detected. A previous study has shown fragmentation-based ML models that had performed well even amongst low coverage sequencing. It showed the ability of ML and AI to detect more than 50 cancer types in one blood test[11]. AI can also reduce some of the background noise and enable reliable detection of ctDNA variants at frequencies as low as 10^{-5}[12] Another reason the applicability of AI is beneficial is its ability to understand the shape of ctDNA. The ctDNA does not consist solely of recognizable nucleotide sequences but also has fragments of different lengths that it can break down into[13]. This study will further delve into the applications of AI and ML in cancer detection through blood based MCED technologies.

2. Materials

2.1. cfDNA methylation

cfDNA is formed by the release of short DNA fragments into the bloodstream through apoptosis of cells. These can be normal cells or tumor cells, which release circulating tumor ctDNA. The ctDNA carries epigenetic alterations that can provide further insight about the cells. One of the major alterations seen is DNA methylation, is the addition of a methyl group to cytosine in CpG dinucleotides that occurs in the promoter or enhancer regions of genes[14,15].

DNA methylation is tightly regulated in normal tissues, leading to stable gene expression, X-inactivation and genomic imprinting. However, in cancer, focal hypermethylation occurs alongside global hypomethylation of CpG islands in regulatory regions of tumor suppressor genes. This leads to oncogenesis through the silencing of negative regulators and subsequent alterations of chromatin structure and gene expression. These abnormal methylation patterns are saved into the ctDNA thus allowing it to be used as a useful biomarker[16-19]. Compared with mutation-based assays, ctDNA methylation markers have a denser signal and can be informative even with low amounts. This is because of the thousands of aberrant CpGs that exist and can be identified easily. They also achieve high specificity and have performed well in early stage MCED testing clinically[20,21].

Recent developments have allowed quantification of methylation at hundreds to thousands of CpG sites from trace amounts of plasma. Some of these methods are targeted bisulfite sequencing, methylation specific PCR and targeted panel assays[16,17]. The strategy being used is identifying the differentially methylated regions (DMRs) and seeing if they are detectable in ctDNA. If they are detectable the machine learning or artificial intelligence models can integrate these loci into a composite score[16,20].

2.2. MCED tests

Multi-Cancer Early Detection tests are a blood-based test that can detect multiple cancers from a single liquid biopsy. These tests provide evidence of a cancer signal[22]. These tests are available through a few sources and largely incorporate the use of cfDNA methylation. The few major companies that have developed these are Galleri, CancerSEEK, PanSeer, EpiPanGI and OneTest.

Galleri is one of the most extensively studied. A previously described study containing over 6000 participants using a MCED test, detected 50 cancer types with a specificity of 99.3% and increasing specificities with stage[11]. Another study conducted by Klein et al. validated Galleri as well. It found that these tests were able to find the cancer signal origin and additional cancers beyond the standard screenings[23].

CancerSEEK uses blood tests to determine cfDNA mutations with eight protein biomarkers. In a study of over 1000 participants completed by Cohen et al. regarding cancer type and staging, reported a median sensitivity of about 70% at a 99% specificity. The test also had the ability to localize the region in the body where the cancer was most likely detected[24].

A study performed by Chen et al. on PanSeer testing, provided evidence that cancer was detected in 91% of individuals who were clinically diagnosed within 4 years of the sample collection at a specificity of about 95%. As this test was performed prior to the actual cancer diagnosis, it is likely that these signals could provide evidence in early detection[25].

EpiPanGI is a pan gastrointestinal multi-cancer methylation panel. In a study performed by Kandilla et al. over 1700 tumor and normal tissue samples from six gastrointestinal cancers were tested and identified with over 60,000 DMRs. They were able to derive site specific localization within the GI tract[26].

OneTest in comparison to the others has been using a different method of detection. Instead of the cfDNA based tests it focuses on existing serum tumor markers with machine learning algorithms. A study containing over 40,000 patients completing their health checkups, found that machine learning outperformed single marker thresholds for predicting cancer[27].

2.3. Biomarkers

Many biomarkers have been found to correlate with different cancers. We will focus on a few that are most reviewed in literature for some of the major cancers. Colorectal cancer is one of the most studied in terms of cfDNA methylation biomarkers. The development of CRC involves promoter hypermethylation in Wnt signaling, DNA repair and cytoskeletal organization[28,29]. Hypermethylation of the promoter SEPT9 was the original basis for the first FDA approved blood-based CRC test. SEPT9 is responsible for encoding a cytoskeletal GTP binding protein that is implicated in cytokinesis and cell shape. The hypermethylation leads to reduced expression and is part of the early process of the adenoma to carcinoma sequence[28-30]. A meta-analysis included 25 studies with about 7000 subjects reported a sensitivity of 67-69% and specificity of 89-92% for the methylated SEPT9 to detect CRC. It was found to have a better sensitivity for stage II-IV cancer development but still had some detectability in stage I cases as well[30,31].

Another research conducted recently focusing on early-stage disease found that optimized algorithms could yield sensitivities of about 80% and specificities of 80-90% for stage I-III cancers using SETP9[28-31]. SDC2 is another CRC related gene that is found to be hypermethylated in tumors. It is a heparan sulfate proteoglycan that has shown greater than 90% methylation frequency in CRC compared with normal mucosa[32] Clinical studies have shown the detection has sensitivities in the 70-90% range alongside specificities greater than 85%[32-34]. It has been found that composite panels including some of these markers together including SDC2 have further improved accuracy. A study completed in 2024 integrating SDC2 SEPT9 and Vimentin (VIM) promoters in stool samples produced sensitivities of 88-92% and specificities of 90-95% for CRC and other advanced adenomas[35]. VIM encodes an intermediate filament protein that is involved with mesenchymal transition and is another feature that is recurrent in CRC[28,36].

Another research done by Borobova, et al showed that a plasma assay with Septin 9 (SEPT9) and Syndecan 2 (SDC2) detected Colorectal Cancer (CRC).as well as high risk precancerous lesions with a high specificity[32]. Although single markers have been useful, further developments of assays focusing on DMRs from multiple genes are being created. A study from Long et al. focused on three Differentially Methylated Regions (DMRs), Disabled Homolog 1 (DAB1), Protein Phosphatase 2 Regulatory Subunit B′ Gamma (PPP2R5C), Family with Sequence Similarity 19 Member A5 (FAM19A5) that had a methylation status in cfDNA that differentiated itself from the controls. It was found that the AUC was 0.76 with a 64% sensitivity and 78% specificity[37]. Zhao et al also developed another blood-based test ColonSecure looking at cancer specific CpG island methylation patterns and reported sensitivities greater than 80% at specificities greater than 90% in detection of stage I-II Colorectal Cancer CRC[38].

Lung Cancer has also been researched recently for cfDNA methylation factors. Research completed by Machado, et al.[38]. pooled results from 44 studies containing about 4000 participants found a sensitivity of 54% and specificity of 86% for markers overall. The significant markers found were RASSF1A, APC, SHOX2, SOX17 and HOXA9[39]. Other studies using Brochoalveolar lavage have shown that SHOX2 and RASSF1A detected cancer at sensitivities of over 80% along with specificities greater than 95% which was even better than conventional methods. [40] SHOX2 and RAFFS1A promoter methylation were used in a study looking at early-stage lung adenocarcinoma using qPCR in blood. As compared to the benign controls the biomarkers distinguished the cancer development with an AUC greater than 0.75 and sensitivities between 75-80% with specificities greater than 90%[41-42]. The data overall empirically supports the importance of the cfDNA biomarkers.

Research on biomarkers of Breast cancer also has some overlap with the others. A study completed by Salta et al. which evaluated a seven gene promoter methylation panel focused on: Adenomatous Polyposis Coli (APC), Breast Cancer Gene 1 (BRCA1), Cyclin D2 (CCND2), Forkhead Box A1 (FOXA1), Phosphoserine Aminotransferase 1 (PSAT1), Ras Association Domain Family Member 1A (RASSF1A), Secretoglobin Family 3A Member 1 (SCGB3A1). Combined methylation of Adenomatous Polyposis Coli (APC), Forkhead Box A1 (FOXA1), Ras Association Domain Family Member

1A (RASSF1A). detected breast cancer at greater than 70% sensitivity and specificity. Models that contained a broader approach with more loci found accuracies greater than 90%[43]. Ras Association Domain Family Member 1A (RASSF1A), Paired-Like Homeodomain Transcription Factor 2 (PITX2). have been shown to be associated with prognosis in breast cancer. A study with more than 300 subjects found Ras Association Domain Family Member 1A (RASSF1A), Paired-Like Homeodomain Transcription Factor 2 (PITX2). to be indicators of poor overall survival from disease independent of the regular clinical variables[44]. Glutathione S-Transferase P1 (GSTP). promoter hypermethylation is one of the consistent alterations seen in prostate cancer[45,46]. Recent data from a meta-analysis confirmed high specificities for Glutathione S-Transferase P1 (GSTP) promoter methylation for prostate cancer diagnosis and has also been included in the panels as a complement to the already existing Prostate-Specific Antigen (PSA). testing[47,48]. More markers exist as this study is just brushing the surface of what can be accomplished with this research. While these biomarkers play different roles in cell signaling, regulation, repair or synthesis they have an interconnected mechanism that can lead to promoter hypermethylation.

3. Methods

According to Sahoo et al. in *Epigenetics & Chromatin*[49], this research paper puts forth multiple hypotheses centered on enhancing diagnostic accuracy, identifying subtle biomarker patterns and facilitating multi-cancer categorization to assess whether Artificial Intelligence (AI), Machine Learning (ML), Large Language Models (LLMs) and Specialized Language Models (SLMs) may improve multi-cancer early detection using bloodstream biomarkers. A wide range of ML classifiers, such as Random Forest (RF), Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), XGBoost-LogitRaw (XGB1), XGBoost-Logistic (XGB2), XGBoost variations, Linear Discriminant Analysis (LDA), Constant-Time Ensemble Learning Classifier (CTELC), Bagging Classifier (BC) and Easy Ensemble Classifier (EE), are applied to the proposed biomarker dataset in order to evaluate these hypotheses.

In healthcare, various AI/ML models and frameworks are currently employed or being researched for Multi-Cancer Early Detection (MCED) using bloodstream data, particularly through liquid biopsy signals such as cfDNA, ctDNA, methylation patterns and other circulating biomarkers.

3.1. Models for multi-cancer early detection (MCED)

3.2.1. DELFI (DNA Evaluation of Fragments for Early Interception): According to Erfan Aref-Eshghi1 et al., a machine learning approach that evaluates cfDNA fragmentation patterns across the genome to diagnose different cancer types with high sensitivity and specificity, as well as predict tissue of origin from blood samples. The model makes it possible to compare the performance of ensemble-driven classification, pattern-discovery depth and linear versus nonlinear learning[50].

3.2.2. MERCURY model: According to the Hajjar, M et al., in the healthcare industry, The MERCURY Model combines numerous cfDNA fragmentation features (such as fragment size distributions, end motifs and copy number variations) with machine learning classifiers to provide robust multi-cancer detection and tissue-of-origin prediction[51].

3.2.3. cfDNA methylation-based classifiers (e.g., Galleri®, PanSeer®, OverC®): According to Irshad. M et al., the Commercial or research MCED platforms that use deep learning or SVM-based models trained on cfDNA methylation signatures to identify cancer signals and, in many cases, infer organ origin[52].

3.2.4. CancerSEEK / Ensemble models: according to Yongjie Xu, et al., Clinical Epigenetics, this model is used in this research to assess several cancers using circulating biomarkers such as proteins and DNA. For instance, machine learning ensembles (typically include logistic regression, random forests or deep learning) to assess several cancers using circulating biomarkers such as proteins and DNA[53].

3.2.5. Elastic net / GLMNET models: According to Irshad M, et al., this model has been employed in cfDNA fragmentation and methylation studies to manage high-dimensional liquid biopsy features while balancing feature selection and classification performance in MCED situations[52].

According to Oyeniyi, J., Oluwaseyi, P, et al., Molecular Cancer studies additionally look into semi-supervised variational autoencoders for specific biomarkers (such as circulating non-coding RNAs) to improve cancer signal detection performance and Table 1 showed the current AI LLM technologies in cancer diagnostics and treatments [54].

Table 1: Current AI technologies in cancer diagnostics and treatment.

AI Tool	Application	AI Model Used	Dataset Used	Performance Benchmak	Clinical Validation Status	Limitation/ Challenges	Data Type Processed	Regulatory Status	Use in Clinical Practice
Deep Variant (Google)	NGS variant calling	Deep learning (CNN)	1000 Genomes, Genome in a bottle	Higher accuracy than traditional metods	Used in research & clinical settings	Requires extensive computational power	Genomic sequencing	Research use only	Limited use in hospitals
Alpha Fold	Protein structure prediction (proteomics)	Transformer-based deep learning	Protein Data Bank (PDB)	RMSD accuracy improvement	Validated on CASP challenges	Limited to known protein sequences	Proteomics	Research use only	Used in pharmaceutical R&D
IBM Watson for Oncology	Precision oncology	NLP, ML-based decision support	Large genomic & clinical datasets	Improved treatment recommendations	Used in select hospitals	Limited explainablity, bias concerns	Clinical & genomic data	Some FDA-Approved components	Used in cancer treatment centers
AI-Driven Liquid Biopsy Analysis	ctDNA & CTC-based cancer detection	ML, CNN- based	Large liquid biopsy datasets	Higher specificity & sensitivity for early detection	Reseach phase, some trails ongoing	Standardization issues in clinical application	Liquid biopsy, ctDNA	Not FDA-approved yet	Limited trails for early detection
CancerSEEK AI	Multi-cancer blood test	ML-based ensemble models	CancerSEEK cohort (10,000 + Patients)	Early detections of 8+ cancers	Clinical trails ongoing	Cost of implementation, false positives	Circulating biomarkers	Not FDA-approved yet	Pilot trails in screening programs

PRS-AI	Personalised risk assesment	AI-based polygenic risk scores (PRSs)	UK biobank, large GWASs	More-accurate risk stratification	Used in some precision medicine initiatives	Ethical concerns, population bias	Genomic risk profiling	Not FDA-approved yet	Early-stage implementation in genetic counselling
AI for Drug Response Prediction	Precision oncology drug matching	ML, Deep reinforcement learning	GDSC, CCLE Datasets	Improved patient-specific therapy recommendations	Limited clinical traits	Requires extensive validation	Genomic & drug response data	Not FDA-approved yet	Limited use in precision oncology

Source: Tiwari, et al. Molecular Cancer research[55].

The table above represents the NGS next generation sequencing, convolutional neural network(CNN), root mean squared deviation(RMSD), critical assessment of protein structure prediction(CASP), research and development(R&D) , natural language processing (NLP), machine learning (ML), Food and Drug Administration(FDA), circulating tumor DNA(ctDNA), circulating tumor cell(CTC), genome-wide association studies (GWASs), Genomics in Drug Sensitivity (GDSC) in Cancer, Cancer Cell Line Encyclopedia(CCLE).

To directly address each hypothesis, several experimental groups are created to ascertain which algorithms are most effective at identifying early cancer signatures and differentiating between cancer types.

a. Experiments where missing values were eliminated.
b. Tests utilizing classifiers that can handle missing-value datasets without eliminating missing values.
c. Imputation techniques are used in experiments to fill in missing values before the LLM-based predictive model is applied **(Figure 1)**.

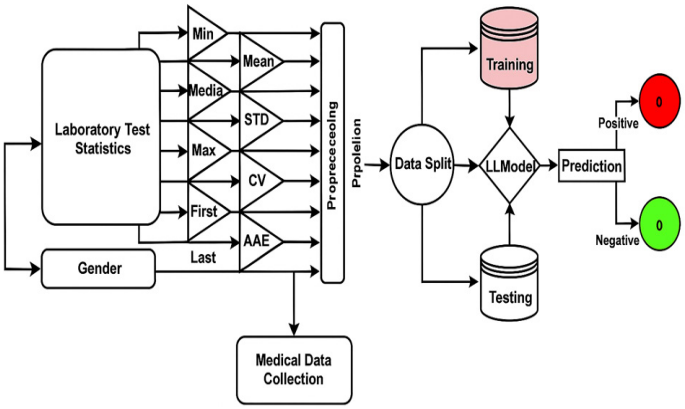

Figure 1: Predictive Model.

A machine learning workflow for binary health prediction utilizing medical data is depicted in the figure, First, last, minimum, median, standard deviation (STD), coefficient of variation (CV), average absolute error (AAE) and other statistical features are derived from patient-specific data, including organ-specific measurements and clinical metrics. After that, these characteristics are sent into a machine learning classifier, which evaluates the information to classify health outcomes as either favorable or negative. The method shows how AI/ML approaches may be used to convert organized medical data into predictive insights.

The following three hypotheses are developed to assess how well artificial intelligence can advance bloodstream biomarker analysis for multi-cancer early detection. These hypotheses seek to evaluate the accuracy of diagnoses, find hidden biomarker trends and ascertain whether Artificial Intelligence (AI), Machine Learning (ML), can identify several types of cancer from a single blood test.

3.1. Hypothesis #1: Diagnostic accuracy

When compared to traditional statistical or single-biomarker techniques, artificial intelligence models trained on bloodstream biomarker profiles would greatly increase the sensitivity and specificity of multi-cancer early detection.

3.2. Hypothesis 2: Early signals and pattern recognition

Compared to existing clinical screening methods, AI-driven feature extraction will discover subtle, non-linear biomarker patterns in blood samples that correspond with early-stage malignancies, allowing identification.

3.3. Hypothesis 3: Multi-cancer classification

Integrating multimodal biomarker data (e.g., cfDNA, proteins, metabolites) into AI-based models will enhance the ability to accurately differentiate between multiple cancer types from a single blood test.

Since conventional screening methods are insufficiently sensitive for early-stage malignancies, this hypothesis was chosen. To test this, we will train AI/ML models on sizable biomarker datasets and use metrics like AUC, sensitivity and specificity to compare their diagnostic performance with current clinical techniques. Early cancer indications in bloodstream biomarkers are weak and can go unnoticed by traditional analytics. Natural Language Processing (NLP), Deep Learning (DL), Convolutional Neural Network (CNN). architectures and LLM-based classifiers are examples of AI feature-extraction approaches that will be utilized to find hidden biomarker patterns and assess their potential correlation with early-stage disease.

Integrating multiple datasets may enhance categorization because individual biomarkers rarely differentiate between different cancer types. This will be tested by creating a multimodal AI model and evaluating how well it can determine the kind and origin of cancer from a single blood sample.

4. Artificial intelligence (AI) and machine learning (ML) in Multi Early Cancer Detection

Artificial intelligence (AI) and machine learning (ML) are currently used in many more cancer diagnosis techniques; non-invasive cancer diagnosis is one of these technologies' biggest benefits.[13], In this case, using data from actual electronic health records (EHR) makes it possible to predict cancer with a very high degree of accuracy.[56] The practical advantages of incorporating basic laboratory tests in the early phases of cancer patient diagnosis are demonstrated by the incorporation of medical data, which may improve predicting. [57] LLM models can be tested and trained on a sizable dataset obtained from medical records. It has been demonstrated that

using LLM, machine learning to complete blood count data (CBC) can improve the identification of cancer. In a related study, decision trees and cross-validation methods were used in conjunction with age, sex and blood count data (CBC) combined with decision trees and cross-validation techniques were employed to accurately predict colorectal cancer.

4.1. Applying LLM to data

Applying the model to data from various demographics yielded an accuracy of 98 to 99%. LLM has been used in various research to identify a variety of cancer types [58]. We employed Artificial Intelligence (AI), Machine Learning (ML), Large Language Model (LLM), with a multi shot learning prompt to identify Colorectal Cancer (CRC). A total of 1035 electronic medical records, comprising both normal and atypical instances, were used in our investigations. With an Area Under the Curve (AUC) of 0.895, Sensitivity (Sens) of 88.5%, Specificity (Spec) of 83.5%, Positive Predictive Value (PPV) of 89.4%, Negative Predictive Value (NPV) of 89.9%, we analyzed satisfactory findings.

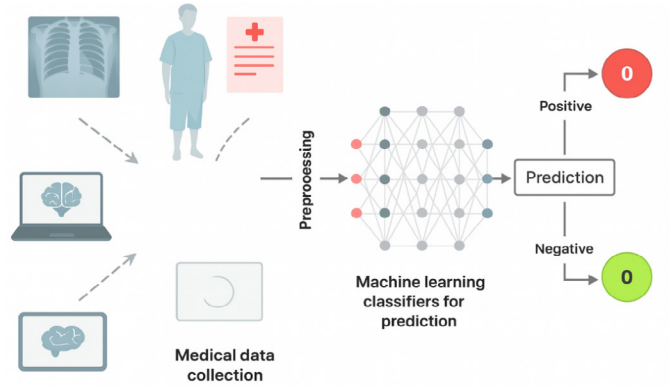

Figure 2: Proposed cancer detection system.

According to Tsai, et al.[59], bladder cancer detection can be enhanced by using clinical laboratory data and machine learning to get around the drawbacks of traditional urine cytology. Using data analyzed from 556 individuals with various cancer kinds, it evaluates five machine learning models using a two-step feature selection procedure. The light gradient boosting machine (lightGBM) model was the most successful, with accuracy rates between 89.8% and 88.9%, sensitivity between 84% and 87.8%, specificities between 82.9% and 86.7% and AUCs between 0.88 and 0.92. This technique demonstrates how ML and clinical data might improve bladder cancer detection.

5. Results and Discussion

Apart from Random Forest (RF), Logistic Regression (LR), Linear Discriminant Analysis (LDA)., where the p-values for these pairs are above the significance level of 0.05, the statistical test results, as shown in **(Table 2)**, demonstrate a substantial difference between SVM and all other approaches. Comparing RF to XGBoost-Logistic (XGB2), K-Nearest Neighbors (KNN), Constant-Time Ensemble Learning Classifier (CTELC), Easy Ensemble Classifier (EEC), Artificial Neural Network (ANN). reveals statistically significant differences as well.

As seen in **(Table 2 and Figure 3)**, Based on Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves, Support Vector Machine (SVM) and Logistic Regression (LR) outperformed Linear Discriminant Analysis (LDA) with an AUC of 0.74. The Easy Ensemble Classifier (EEC) had the lowest performance, with an AUC of 0.34.

All these models' measures, including the AUC values, were insufficient for clinical application. This holds true even for the finest model that was looked at. We used principal component analysis (PCA) and IsoMap to visualize the suggested dataset in lower dimensions to investigate the possible cause of these results. The projection of the suggested dataset on the first two elements of the two projection techniques employed is shown in **(Figure 2)**.

This explains why certain models perform poorly. The classifier finds it challenging to differentiate between the classes because they overlap. Examining the characteristics that contribute to the classification choice (cancer or not) is crucial in the medical industry. To identify the most crucial features for making predictions, we use permutation feature importance (PFI). The RF classifier is used to accomplish this for two primary purposes. The first is that RF has shown suitability for PFI29 and the second is that RF already offers high precision on the suggested dataset. The most crucial characteristics are shown in **(Figure 1)**, which serves as the basis for the decision prediction.

Table 2: Wilcox Signed Rank Test p-Values.

	SVM	RF	LR	ANN	XGB1	XGB2
SVM	1	0.784	0.784	0.784	0.784	0.784
RF	0.177	1	0.77	0.77	0.77	0.77
LR	0.081	0.582	1	0.76	0.76	0.76
ANN	0.003	0.003	0.036	1	0.759	0.737
XGB1	0.032	0.175	1.029	0.004	1	0.759
XGB2	0.007	0.031	0.228	0.086	0.013	1
KNN1	0.001	0.004	0.008	0.006	0.002	0.004
KNN3	0.004	0.009	0.009	0.002	0.009	0.009
BC	0.002	0.004	0.004	0.004	0.004	0.004
CTELC	0.002	0.004	0.034	0.004	0.004	0.004
EEC	0.001	0.008	0.619	0.004	0.219	0.001
LDA	0.083	0.597	1.055	0.037	1.04	0.262

The Wilcoxon Signed-Rank Test p-Values, which were computed for the results and compared the performance of each pair of classifiers, are displayed in the bottom triangle of the table. The best-performing classifier in terms of F1-score is indicated by the top half of the table, which shows the maximum mean precision for each classifier pair.

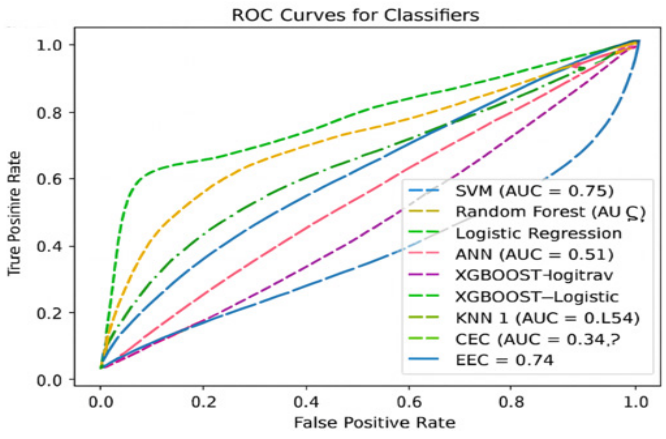

Figure 3: ROC Curves for Classifiers.

While this study highlights the potential of AI in MCED, several considerations must be addressed prior to clinical translation. The use of minimally invasive, blood-based biomarkers in combination with AI-driven analytical methods offers a scalable and patient-centered approach that may enhance early detection, particularly among asymptomatic individuals and populations with low adherence to conventional screening strategies.

However, beyond model performance, important challenges remain in real-world implementation. Integration into clinical practice requires standardization of biomarker collection and processing protocols, as well as compatibility with electronic health record systems. In addition, successful adoption will depend on the interpretability and transparency of model outputs to support clinician confidence in decision-making.

Another key limitation is the need for broader validation across diverse patient populations. Variability in demographic characteristics, comorbid conditions and environmental exposures may influence biomarker expression and model performance, underscoring the importance of ensuring generalizability across heterogeneous cohorts.

Furthermore, ethical and regulatory considerations, including data privacy, potential algorithmic bias and approval processes, must be carefully addressed prior to widespread implementation. Economic feasibility also remains an important factor, as the integration of MCED systems requires investment in infrastructure, validation studies and cost-effectiveness evaluation.

Overall, while the findings are promising, successful clinical integration of AI-based MCED systems will depend on addressing these practical, regulatory and system-level considerations.

6. Conclusion

This research paper demonstrates that Artificial Intelligence (AI), Machine Learning (ML), Large Language Models (LLMs) and Specialized Language Models (SLMs) can significantly enhance multi-cancer early detection by improving diagnostic accuracy, discovering subtle biomarker patterns and enabling robust multi-cancer classification from bloodstream data. We also reviewed various ML classifiers demonstrating the efficacy of ensemble and deep-learning techniques in collecting complicated cfDNA fragmentation, methylation and circulating biomarker signals. The results are compatible with known MCED frameworks such as DELFI, MERCURY, Galleri®, PanSeer® and CancerSEEK, confirming the clinical validity of AI-driven liquid biopsy techniques. This study allows the integration of advanced AI models into scalable MCED systems, allowing for earlier diagnosis, more informed clinical decision-making and better patient outcomes.

There is a great deal of potential to build on the predictive modeling strategies employed in this study by utilizing more sophisticated approaches in further research. Investigating deep learning techniques may reveal more intricate patterns in the information, which could enhance the precision and resilience of the forecasts[60,61].

Experimenting various imputation methods to handle missing values in the dataset is another way to improve it. Despite their effectiveness, current approaches have drawbacks that new imputation techniques could address, improving the dataset's overall quality and the model's performance.

By using 5-fold cross-validation to evaluate unseen data, the current model performs well, guaranteeing both training and testing of all the data. Nevertheless, the highest diagnosis accuracy attained is just 72%, which might be adequate for some uses but is insufficient for medical ones. When compared to medical professionals, our results. This discrepancy results from our model's exclusive reliance on blood tests and demographic variables like age and sex, while medical practitioners employ other data like tumor markers, imaging and scans to improve their diagnoses. We might greatly increase the ML model's diagnostic accuracy by including this extra data since more thorough data would allow it to make more accurate judgments. We can make such advancements in our upcoming work.

7. References

1. Siegel RL, Miller KD, Wagle NS, et al. Cancer statistics, 2025. CA Cancer J Clin, 2025;75: 10-45.
2. Mannelli C. Tissue vs liquid biopsies for cancer detection: ethical issues. J Bioeth Inq, 2019;16(4): 551-557.
3. Connal S, Cameron JM, Sala A, et al. Liquid biopsies: the future of cancer early detection. J Transl Med, 2023;21: 118.
4. Mariotto AB, Enewold L, Zhao J, et al. Medical care costs associated with cancer survivorship in the United States. Cancer Epidemiol Biomarkers Prev, 2020;29(7): 1304-1312.
5. Laudicella M, Walsh B, Burns E, et al. Cost of care for cancer patients in England: evidence from population-based patient-level data. Br J Cancer, 2016;114(11): 1286-1292.
6. Geneve N, Kairys D, Bean B, et al. Colorectal cancer screening. Prim Care Clin Office Pract, 2019;46(1): 135-148.
7. Sedani AE, Gomez SL, Lawrence WR, et al. Social risks and nonadherence to recommended cancer screening among US adults. JAMA Netw Open, 2025;8(1): 2449556.
8. Marrugo-Ramírez J, Mir M, Samitier J. Blood-based cancer biomarkers in liquid biopsy: a promising non-invasive alternative to tissue biopsy. Int J Mol Sci, 2018;19(10): 2877.
9. Wan JCM, Massie C, Garcia-Corbacho J, et al. Liquid biopsy: from discovery to clinical application. Cancer Discov, 2021;11(4): 858-873.
10. Ahlquist DA. Universal cancer screening: revolutionary, rational and realizable. NPJ Precis Oncol, 2018;2: 23.
11. Klein EA, Richards D, Cohn A, et al. Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. Ann Oncol, 2021;32(9): 1167-1177.
12. Kuligina ES, Yanus GA, Imyanitov EN. Improvement of the sensitivity of circulating tumor DNA-based liquid biopsy: current approaches and future perspectives. Explor Target Antitumor Ther, 2025;6: 1002333.
13. Cristiano S, Leal A, Phallen J, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature, 2019;570: 385-389.
14. Laird PW. The power and the promise of DNA methylation markers. Nat Rev Cancer, 2003;3(4): 253-266.
15. Rendek T, et al. Current challenges of methylation-based liquid biopsies in oncology. Cancers (Basel), 2024;16(11): 2001.
16. Medina JE, et al. Cell-free DNA approaches for cancer early detection and interception. J Immunother Cancer, 2023;11: e006013.
17. Ibrahim J, Peeters M, Van Camp G, et al. Methylation biomarkers for early cancer detection and diagnosis: current and future perspectives. Eur J Cancer, 2023;181: 1-17.
18. Pharo H, et al. A roadmap for DNA methylation biomarkers in liquid biopsies. Oncogene, 2025;44.

19. Li L, et al. Circulating tumor DNA methylation detection as biomarker and its application in tumor liquid biopsy: advances and challenges. MedComm, 2024;5: 766.
20. Oliver J, et al. Emerging noninvasive methylation biomarkers of cancer prognosis and drug response prediction. Semin Cancer Biol, 2022;83: 584-595.
21. Johnston AD, et al. Epigenetic liquid biopsies for minimal residual disease detection in solid tumors. Front Oncol, 2023;13: 1103797.
22. Cheung LC, Katki HA, Pinsky PF, et al. multi-cancer early detection: the new frontier in cancer early detection. Annu Rev Med, 2025;76.
23. Klein EA, Richards D, Cohn A, et al. Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. Ann Oncol, 2021;32(9): 1167-1177.
24. Cohen JD, Li L, Wang Y, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science, 2018;359(6378): 926-930.
25. Chen X, Gole J, Gore A, et al. Noninvasive early detection of cancer four years before conventional diagnosis using a blood test. Nat Commun, 2020;11: 3475.
26. Kandimalla R, et al. EpiPanGI Dx: a cell-free DNA methylation fingerprint for the early detection of gastrointestinal cancers. Clin Cancer Res, 2021;27(22): 6135-6144.
27. Wang HY, Chen CH, Shi S, et al. Improving multi-tumor biomarker health check-up tests with machine-learning algorithms. Cancers (Basel), 2020;12(6): 1442.
28. Fatemi N, et al. DNA methylation biomarkers in colorectal cancer: clinical applications for precision medicine. Int J Cancer, 2022;151: 2068-2081.
29. Wang Y, et al. Research progress of DNA methylation in colorectal cancer. Mol Med Rep, 2024;29.
30. Song L, et al. Performance of the SEPT9 gene methylation assay in colorectal cancer screening: a meta-analysis. Sci Rep, 2017;7: 3032.
31. Hariharan R, Jenkins M. Utility of the methylated SEPT9 test for the early detection of colorectal cancer: a systematic review and meta-analysis. BMJ Open Gastroenterol, 2020;7: 000355.
32. Borobova V, Aksamentov A, Sazonov D, et al. Analysis of SDC2 and SEPT9 promotes methylation in plasma cfDNA to detect colorectal and precancerous lesions. Explor Med, 2025;6: 1001322.
33. Oh TJ, et al. Genome-wide identification and validation of a novel methylation biomarker, SDC2, for blood-based detection of colorectal cancer. J Mol Diagn, 2013;15: 498-507.
34. Zhao G, et al. Multiplex methylated DNA testing in plasma with high sensitivity and specificity for colorectal cancer. Cancer Med, 2019;8: 5619-5628.
35. Liu Y, et al. DNA methylation analysis of SDC2, SEPT9 and VIM in colorectal cancer diagnosis. Cancer Med, 2024
36. Oh CK, et al. Pathogenesis and biomarkers of colorectal cancer by liquid biopsy. Intest Res, 2024
37. Long Z, et al. Discovery and validation of methylation signatures in cell-free DNA for colorectal cancer detection. Biomolecules, 2024;14: 996.
38. Zhao F, et al. Efficacy of a cell-free DNA methylation-based blood test (ColonSecure) for colorectal cancer screening. Mol Cancer, 2023;22: 157.
39. Machado DI, et al. Circulating cell-free DNA methylation as biomarker for lung cancer diagnosis: a systematic review and meta-analysis. Cancers (Basel), 2025
40. Zhang C. et al. DNA methylation analysis of the SHOX2 and RASSF1A panel in bronchoalveolar lavage fluid for lung cancer diagnosis. J Cancer, 2017;8: 3585-3592.
41. Jin Y, et al. DNA methylation analysis in plasma for early diagnosis of lung adenocarcinoma based on SHOX2 and RASSF1A methylation. Medicine (Baltimore), 2024;103: 40042.
42. Chen Y, et al. SHOX2 and RASSF1A methylation for early-stage lung adenocarcinoma diagnosis. Mol Clin Oncol, 2025
43. Salta S, et al. A DNA methylation-based test for breast cancer detection in circulating cell-free DNA. J Clin Med, 2018;7: 420.
44. Schwarzenbach H, Hoon DSB, Pantel K. Cell-free nucleic acids as biomarkers in cancer patients. Breast Cancer Res, 2015;17: 136.
45. Wu T, et al. Measurement of GSTP1 promoter methylation in body fluids: a biomarker for prostate cancer diagnosis. Br J Cancer, 2011;104.
46. Mair R, et al. Cell-free DNA technologies for the analysis of brain cancer. Br J Cancer, 2022;126: 375-387.
47. Ye J, et al. Glutathione-S-transferase P1 promoter methylation in cfDNA for prostate cancer diagnosis: a meta-analysis. Prostate Cancer Prostatic Dis, 2023;26.
48. He W, et al. Cell-free DNA in the management of prostate cancer. Transl Androl Urol, 2023;12.
49. Sahoo K, Lingasamy P, Khatun M, et al. Artificial Intelligence in cancer epigenomics: a review on advances in pan-cancer detection and precision medicine.
50. ArefEshghi E, Abadi AB, Farhadieh ME, et al. DNA methylation and machine learning: challenges and perspective toward enhanced clinical diagnostics
51. Hajjar M, Albaradei S, Aldabbagh G. Machine Learning Approaches in Multi-Cancer Early Detection. Information, 2024;15: 627.
52. Irshad M, Hanif MI, Khan M, et al. Revolutionizing Healthcare Delivery: Evaluating the Impact of Google's Gemini AI as a Virtual Doctor in Medical Services. J Artif Intell Mach Learn & Data Sci, 2024;2(2): 1618-1625.
53. Xu Y, Zhu S, Xia C, et al. Liquid biopsy-based multi-cancer early detection: an exploration road from evidence to implementation, 2025;70(17): 2852-2867.
54. Oyeniyi J, Oluwaseyi P. Emerging trends in AI-powered medical imaging: enhancing diagnostic accuracy and treatment decisions. J Int J Enhanced Res Sci Technol Eng, 2024;13: 2319-7463.
55. Tiwari A, Mishra S, Kuo TR. Current AI technologies in cancer diagnostics and treatment. Mol Cancer, 2025;24: 159.
56. Bandyopadhyay A, Albashayreh A, Zeinali N, et al. Using real-world electronic health record data to predict the development of 12 cancer-related symptoms in the context of multimorbidity. JAMIA Open, 2024;7(3): 082.
57. Al-Khlifeh EM, Alkhazi IS, Alrowaily MA, et al. Extended spectrum beta-lactamase bacteria and multidrug resistance in Jordan are predicted using a new machine-learning system. IDR, 2024;17: 3225-3240.
58. Li H, Lin J, Xiao Y, et al. Colorectal cancer detected by machine learning models using conventional laboratory test data. Technol Cancer Res Treat, 2021;20: 15330338211058352.
59. Tsai IJ, Shen WC, Lee CL, et al. Machine learning in prediction of bladder cancer on clinical laboratory data. Diagnostics, 2022;12(1): 203.
60. Kumar R, Saha P. A review on artificial intelligence and machine learning to improve cancer management and drug discovery. Int J Res Applied Sci Bio, 2022;9(3): 149-156.
61. Tarawneh AS, Celik C, Hassanat AB, et al. Detailed investigation of deep features with sparse representation and dimensionality reduction in CBIR: a comparative study. Intell Data Anal, 2020;24(1): 47-68.