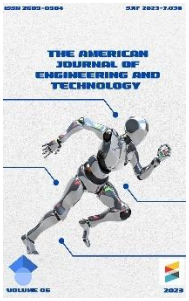
The American Journal of Engineering and Technology
(ISSN – 2689-0984)
VOLUME05ISSUE12Pages:68-87
SJIF IMPACT FACTOR (2020: 5. 32)(2021: 5. 705)(2022: 6. 456) (2023: 7. 038)
OCLC– 1121105677
Publisher:The USA Journals



Journal Website:https://theamericanjournals.com/index.php/tajet

Khamidova, S. (2023). Machine Learning Based Analysis Of Clinical And Laboratory Biomarkers For Early Cancer Detection. The American Journal of Engineering and Technology, 5(12), 68–87. Retrieved from https://theamericanjournals.com/index.php/tajet/article/view/8516

Copyright:Original content from this work may be used under the terms of the creative commons attributes 4.0 licence.

Research Article

Machine Learning Based Analysis Of Clinical And Laboratory Biomarkers For Early Cancer Detection

Submission Date:November25, 2023, Accepted Date: December22, 2023,
Published Date:December29, 2023|

Shakhnoza Khamidova
Independent Researcher,USA

ABSTRACT

Early detection remains an important objective in cancer diagnostics because clinical outcomes are strongly influenced by the stage at which disease is identified. Clinical and laboratory biomarkers provide measurable information about physiological, biochemical, and disease-related changes that may support earlier detection; however, the simultaneous analysis of multiple biomarkers and patient characteristics can be challenging using conventional statistical approaches. Machine learning (ML) offers computational methods for identifying complex patterns within clinical and laboratory datasets and for developing predictive models that may assist in distinguishing cancer from non-cancer conditions. This article examines the application of ML-based analysis to clinical and laboratory biomarkers for early cancer detection, with particular attention to data preparation, feature selection, classification, prediction, and model evaluation. Research involving commonly used clinical measurements, biochemical markers, hematological parameters, tumor-associated biomarkers, and combinations of laboratory and clinical variables is discussed. Supervised learning methods, including logistic regression, decision trees, random forests, support vector machines, k-nearest neighbors, and artificial neural networks, are considered in relation to their use in biomarker-based cancer prediction. The discussion further examines how feature selection and multivariable modelling can identify combinations of biomarkers that provide greater diagnostic information than individual measurements. Findings reported across the literature indicate that ML can improve the analysis of heterogeneous biomarker profiles by capturing nonlinear relationships and interactions among clinical variables. Nevertheless, model performance is

The American Journal of Engineering and Technology
(ISSN – 2689-0984)
VOLUME05ISSUE12Pages:68-87
SJIF IMPACT FACTOR (2020: 5. 32)(2021: 5. 705)(2022: 6. 456) (2023: 7. 038)
OCLC– 1121105677
Publisher:The USA Journals

influenced by dataset size, biomarker selection, data quality, class imbalance, validation strategies, and differences between patient populations. The article concludes that ML-based analysis provides a useful framework for combining clinical and laboratory biomarkers in early cancer detection, while emphasizing the need for independent validation, interpretable models, and clinically representative datasets before widespread diagnostic application.

KEYWORDS

Cancer Biomarkers, cancer detection, early cancer detection, machine learning.

INTRODUCTION

Cancer remains a major global health burden, and the clinical consequences of delayed diagnosis are particularly important because treatment options and outcomes are strongly influenced by the stage at which malignancy is identified. Early detection therefore represents an important component of cancer control, with screening intended to identify disease or precancerous abnormalities before clinical symptoms become apparent, while early diagnosis focuses on identifying symptomatic disease at the earliest possible stage. The distinction is clinically relevant because screening is appropriate only for selected cancer types and populations, whereas timely diagnosis of symptomatic disease remains important across a much broader range of malignancies. The World Health Organization has emphasized that early diagnosis can improve the probability of successful treatment while reducing delays, treatment-related morbidity, and healthcare costs [1].

Despite advances in diagnostic oncology, early cancer detection remains challenging. Many malignancies produce nonspecific clinical manifestations during their initial development, and potentially informative biological changes may be subtle when individual laboratory measurements are considered independently. Conventional diagnostic workflows consequently rely on combinations of clinical assessment, laboratory testing, imaging, pathological examination, and, where appropriate, tumor-specific biomarkers. Although these approaches remain fundamental to cancer diagnosis, the interpretation of multiple measurements simultaneously can be difficult because cancer-associated changes may be distributed across several physiological systems rather than being represented by a single highly specific marker. This creates an opportunity for computational approaches capable of identifying multidimensional patterns within routinely collected clinical information.

Clinical and laboratory data are particularly attractive for this purpose because they are generated routinely during patient assessment. Clinical variables such as age, sex, smoking history, family history, symptoms, and comorbidities can provide information about baseline cancer risk, whereas laboratory investigations can capture systemic changes associated with malignancy. Hematological parameters, for example, may reflect anemia, inflammation, immune responses, or alterations in circulating blood-cell populations. Biochemical measurements can provide information regarding hepatic, renal, metabolic, or nutritional status, while established tumor-associated markers may contribute more disease-specific information. Rather than treating these measurements as independent diagnostic entities, machine learning (ML) provides a framework for evaluating their joint predictive contribution.

An important early demonstration of this concept was provided by Hornbrook et al. [2], who evaluated a machine-learning model for colorectal cancer (CRC)

The American Journal of Engineering and Technology
(ISSN – 2689-0984)
VOLUME05ISSUE12Pages:68-87
SJIF IMPACT FACTOR (2020: 5. 32)(2021: 5. 705)(2022: 6. 456) (2023: 7. 038)
OCLC– 1121105677
Publisher:The USA Journals

detection using gender, age, and complete blood count (CBC) information. Their validation study included 900 patients with CRC and 9,108 controls from the Kaiser Permanente Northwest population. Instead of depending on a dedicated molecular cancer marker, the model extracted predictive information from routinely available CBC measurements together with demographic variables. The reported area under the receiver operating characteristic curve (AUROC) was approximately 0.80, and the model was capable of identifying patients at increased risk of undiagnosed CRC before conventional clinical diagnosis. Importantly, the study demonstrated that routinely generated hematological information could contain a detectable signal associated with malignancy when evaluated collectively rather than through isolated laboratory thresholds

A similar principle was subsequently investigated for lung cancer by Wu et al. [3], who developed a Random Forest model using 49 routine blood indices from 277 individuals, including 183 patients with lung cancer and 94 individuals without lung cancer. Feature analysis identified a combination of 19 routine blood indices associated with lung cancer. The investigators reported cross-validation sensitivity of 96.3%, specificity of 94.97%, and overall accuracy of 95.7%. Of particular relevance to early diagnostic research, the study examined whether routine laboratory measurements rather than specialized molecular assays could contain sufficient collective information to distinguish lung cancer from other clinical conditions, including tuberculosis. The findings provided an empirical basis for investigating ML as a means of extracting clinically meaningful patterns from conventional laboratory testing.

These findings illustrate an important conceptual shift in biomarker-based cancer detection. Traditional biomarker research often seeks a variable with a strong individual association with a particular malignancy. Machine learning, in contrast, can model a feature vector containing multiple variables:

$$\mathbf{X} = [x_1, x_2, \ldots, x_p]$$

where x_i represents an individual clinical or laboratory measurement and p represents the total number of available features. The potential value of this approach is not restricted to one cancer type. Research available before 2023 had already investigated ML-based analysis of routine laboratory measurements in several malignancies and clinical settings. For example, ML approaches had been evaluated for detecting multiple myeloma using routine blood tests, with feature selection incorporated to identify informative laboratory variables and with internal and external validation used to examine model robustness.

However, high predictive performance in an individual study should not automatically be interpreted as evidence of clinical readiness. Laboratory values are influenced by age, sex, inflammation, infection, medication use, comorbidities, laboratory methodology, and population characteristics. Furthermore, retrospective case-control datasets may differ substantially from the heterogeneous populations encountered during real-world screening or primary-care assessment. Consequently, the development of clinically useful ML-based biomarker models requires careful preprocessing, feature selection, avoidance of data leakage, appropriate cross-validation, independent testing, and preferably external validation in populations distinct from the development cohort [4]. Against this background, machine learning-based analysis of clinical and laboratory biomarkers represents a potentially useful strategy for extracting complementary diagnostic information from routinely available patient data. This article therefore examines the application of machine learning to clinical and laboratory biomarkers for early cancer detection. It considers the principal categories of

The American Journal of Engineering and Technology
(ISSN – 2689-0984)
VOLUME05ISSUE12Pages:68-87
SJIF IMPACT FACTOR (2020: 5. 32)(2021: 5. 705)(2022: 6. 456) (2023: 7. 038)
OCLC– 1121105677
Publisher:The USA Journals

clinical and laboratory variables used in predictive models, methods for preprocessing and selecting informative features, commonly applied machine-learning algorithms, and evidence from cancer-specific applications.

Clinical and Laboratory Biomarkers in Early Cancer Detection

Clinical and laboratory biomarkers provide complementary information for cancer detection. Clinical variables describe characteristics and exposures associated with an individual's underlying risk, whereas laboratory measurements provide quantitative information about physiological, hematological, biochemical, and tumor-associated changes. Their diagnostic value may be limited when individual variables are interpreted in isolation; however, their combined analysis can reveal patterns that are difficult to identify using conventional threshold-based approaches. This makes multimarker datasets particularly suitable for machine learning models, which can estimate the joint contribution of numerous potentially correlated features.

Clinical biomarkers and patient characteristics

Demographic and clinical characteristics constitute an important component of cancer prediction models. Age, sex, smoking history, family history, body mass index, previous disease, medication use, and presenting symptoms can influence both the probability of malignancy and the interpretation of laboratory abnormalities. These variables therefore provide contextual information that can complement laboratory measurements.

Age is particularly relevant because cancer incidence generally increases with advancing age, while sex can influence the incidence of specific malignancies and the distribution of several laboratory parameters. Lifestyle and exposure variables can provide additional information; smoking history, for example, is strongly associated with lung cancer risk and can potentially improve discrimination when combined with blood-based measurements. Importantly, these variables are generally risk-associated features rather than cancer-specific biomarkers. Their value in an ML model arises from their contribution to a multivariable pattern rather than from their ability to independently establish a diagnosis.

Clinical information can also help distinguish malignant disease from benign conditions. This distinction is particularly important for cancers in which symptoms or laboratory abnormalities may have multiple possible causes. For example, ovarian cancer can present with findings that overlap with benign ovarian disease. In a study of 349 patients with ovarian tumors, Sönmez et al. evaluated several clinical and laboratory variables using multiple ML algorithms, including Random Forest, SVM, Logistic Regression, XGBoost, Gradient Boosting, and LightGBM. Age and other clinical characteristics were analyzed together with serum tumor markers and routine laboratory measurements, illustrating the potential value of combining patient-level characteristics with biological measurements rather than treating each data type independently [5].

Thus, a clinical feature vector can be represented as:

$$\mathbf{C} = [c_1, c_2, \ldots, c_m]$$

where c_i represents an individual demographic, exposure, symptom, or clinical characteristic. When combined with laboratory measurements L, the complete input to an ML model becomes:

$$\mathbf{X} = [\mathbf{C}, \mathbf{L}]$$

The objective is then to determine whether the combined feature space provides greater discriminatory information than either clinical or laboratory variables considered independently.

The American Journal of Engineering and Technology
(ISSN – 2689-0984)
VOLUME05ISSUE12Pages:68-87
SJIF IMPACT FACTOR (2020: 5. 32)(2021: 5. 705)(2022: 6. 456) (2023: 7. 038)
OCLC– 1121105677
Publisher:The USA Journals

Hematological Biomarkers

Hematological parameters are among the most accessible biomarkers for computational cancer detection because complete blood counts are routinely obtained in many clinical settings. Measurements such as hemoglobin, erythrocyte indices, leukocyte count, neutrophil count, lymphocyte count, monocyte count, platelet count, and related ratios can reflect systemic physiological changes associated with malignancy. Anemia is particularly relevant because chronic blood loss, nutritional disturbances, inflammation, altered iron metabolism, and other cancer-associated mechanisms can affect erythrocyte parameters. Leukocyte and differential counts may provide information about systemic inflammatory and immune responses, while platelet abnormalities can occur through interactions between malignant processes, inflammation, and coagulation pathways. Rather than selecting one abnormal blood parameter as a diagnostic indicator, ML permits simultaneous analysis of multiple hematological measurements. This approach was demonstrated by Hornbrook et al., whose colorectal cancer model incorporated CBC parameters together with age and gender. Their study included 900 patients with CRC and 9,108 controls, and the model achieved an AUROC of approximately 0.80 in the validation population [2]. The importance of this finding lies not simply in the reported performance but in the demonstration that routine hematological information can contain a detectable multivariable signal associated with colorectal cancer.

Derived hematological indices can also provide a compact representation of inflammatory status. Common examples include the neutrophil-to-lymphocyte ratio (NLR) and platelet-to-lymphocyte ratio (PLR):

$$NLR = \frac{\text{Neutrophil count}}{\text{Lymphocyte count}}$$

$$PLR = \frac{\text{Platelet count}}{\text{Lymphocyte count}}$$

Such composite variables may allow ML models to capture relationships between different cellular components more directly than when each count is considered independently.

Biochemical and Metabolic Biomarkers

Routine biochemical measurements provide another important source of features for cancer prediction. Liver-associated enzymes, bilirubin, albumin, creatinine, glucose, calcium, uric acid, and other metabolic variables can reflect alterations in organ function, metabolism, inflammation, nutritional status, or systemic effects of malignancy. Their potential value is particularly apparent when several biochemical measurements are analyzed together. Individual abnormalities often have limited specificity: elevated alanine aminotransferase, for example, can arise from numerous hepatic and non-hepatic conditions and should not be interpreted as an independent indicator of malignancy. ML models can instead evaluate the combined pattern of several biochemical measurements alongside demographic and hematological variables. The ovarian cancer study by Sönmez et al. provides a useful example. In their analysis, laboratory variables identified as informative included alanine aminotransferase, calcium, indirect bilirubin, uric acid, and sodium, alongside hematological parameters and tumor-associated markers. The study evaluated several classification approaches and reported that ML models could distinguish malignant from benign ovarian tumors with accuracy reaching approximately 91% [6].

The American Journal of Engineering and Technology
(ISSN – 2689-0984)
VOLUME05ISSUE12Pages:68-87
SJIF IMPACT FACTOR (2020: 5. 32)(2021: 5. 705)(2022: 6. 456) (2023: 7. 038)
OCLC– 1121105677
Publisher:The USA Journals

Inflammatory and Immune-Related Biomarkers

Inflammation is closely associated with tumor development and progression, and inflammatory changes can be reflected through both conventional laboratory measurements and derived indices. C-reactive protein, erythrocyte sedimentation rate, leukocyte populations, neutrophil-to-lymphocyte ratio, platelet-to-lymphocyte ratio, and other inflammation-related measurements have consequently been investigated as potential components of cancer prediction models. The difficulty is that systemic inflammation is not specific to malignancy. Infection, autoimmune disease, tissue injury, and other inflammatory conditions can produce similar laboratory abnormalities. Consequently, the principal advantage of ML is not that it converts an inflammatory marker into a cancer-specific test, but that it can evaluate inflammatory measurements in conjunction with demographic, clinical, hematological, biochemical, and tumor-associated features. This distinction is important when interpreting ML-derived feature importance. A feature may contribute substantially to classification because it improves separation between the populations represented in a particular dataset without necessarily representing a causal mechanism of carcinogenesis. Consequently, associations identified by predictive models require biological interpretation and validation before individual variables can be considered clinically meaningful biomarkers.

Conventional Tumor-Associated Biomarkers

Conventional tumor-associated biomarkers provide more disease-oriented information than many routine laboratory measurements. Examples include prostate-specific antigen (PSA) for prostate disease, carcinoembryonic antigen (CEA) for gastrointestinal malignancies, carbohydrate antigen 125 (CA-125) for ovarian cancer, carbohydrate antigen 19-9 (CA 19-9) for pancreatic and biliary malignancies, and alpha-fetoprotein (AFP) in hepatocellular carcinoma.

Despite their clinical utility in selected settings, many tumor markers have limitations related to sensitivity, specificity, biological variability, or elevation in nonmalignant conditions. ML can address part of this limitation by combining tumor markers with other patient characteristics and laboratory measurements. The ovarian cancer study by **Sönmez et al.** is a clear example. CA-125, CA 19-9, CEA, and human epididymis protein 4 (HE4) were among the variables identified as important in differentiating malignant from benign ovarian tumors, while routine hematological and biochemical measurements supplied additional information.

The underlying principle can be expressed as:

$$P(Y=1|\mathbf{X}) = P(Y=1|C_1,\ldots,C_m,L_1,\ldots,L_n,T_1,\ldots,T_k)$$

where C represents clinical variables, L represents routine laboratory variables, and T represents tumor-associated biomarkers. Rather than asking whether a single tumor marker exceeds a predefined threshold, the model estimates the probability of malignancy from the **joint distribution of heterogeneous features**.

Major Clinical and Laboratory Biomarker Categories Relevant to Machine Learning-Based Cancer Detection

Biomarker category	Representative variables	Potential information provided	Major limitation
Demographic	Age, sex	Baseline cancer risk and population differences	Low specificity

The American Journal of Engineering and Technology
(ISSN – 2689-0984)
VOLUME05ISSUE12Pages:68-87
SJIF IMPACT FACTOR (2020: 5. 32)(2021: 5. 705)(2022: 6. 456) (2023: 7. 038)
OCLC– 1121105677
Publisher:The USA Journals

Clinical	Symptoms, smoking history, family history, comorbidities	Patient-level risk and clinical context	Subjective/incomplete recording
Hematological	Hemoglobin, WBC, neutrophils, lymphocytes, platelets	Anemia, inflammation, immune and systemic responses	Strong influence from non-cancer conditions
Derived hematological indices	NLR, PLR and related ratios	Composite inflammatory information	Non-specific
Biochemical	ALT, AST, bilirubin, albumin, creatinine, glucose, calcium	Organ function and systemic/metabolic changes	Frequently affected by nonmalignant disease
Inflammatory	CRP, ESR, leukocyte-derived measures	Systemic inflammatory response	Limited cancer specificity
Tumor-associated	PSA, CEA, CA-125, CA 19-9, AFP, HE4	More cancer-oriented biological information	Variable sensitivity and specificity
Multimarker combinations	Clinical + laboratory + tumor markers	Integrated patient-specific predictive pattern	Greater complexity and risk of overfitting

The evidence therefore supports a shift from single-biomarker interpretation toward integrated biomarker profiles. This provides the biological and clinical basis for the next stage of the analytical process: converting heterogeneous laboratory and clinical measurements into a structured dataset suitable for machine learning through preprocessing, feature selection, and model development.

Machine Learning Framework for Biomarker-Based Cancer Analysis

The application of machine learning to clinical and laboratory biomarkers involves a sequence of analytical steps extending from dataset construction and preprocessing to feature selection, model development, and validation. This workflow is particularly important for laboratory-based cancer prediction because clinical datasets frequently contain heterogeneous variables measured on different scales, missing observations, correlated biomarkers, and substantial differences between cancer and non-cancer populations. Consequently, predictive performance depends not only on the choice of classification algorithm but also on how the underlying data are prepared and how model performance is evaluated.

Data Collection and Preprocessing

Clinical and laboratory ML models generally require a clearly defined outcome variable and a corresponding

The American Journal of Engineering and Technology
(ISSN – 2689-0984)
VOLUME05ISSUE12Pages:68-87
SJIF IMPACT FACTOR (2020: 5. 32)(2021: 5. 705)(2022: 6. 456) (2023: 7. 038)
OCLC– 1121105677
Publisher:The USA Journals

set of predictor variables. For cancer detection, the outcome may be represented as a binary variable:

$$Y = \begin{cases} 1, & \text{cancer present} \\ 0, & \text{cancer absent} \end{cases}$$

while the predictor matrix contains clinical characteristics and laboratory measurements:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

where n denotes the number of patients and p the number of candidate features.

The quality of this dataset is fundamental because a model learns the relationships represented in the development population. Gould et al. demonstrated the potential of large clinical datasets for this purpose in a study of early lung cancer identification. Their analysis included 6,505 patients with non-small-cell lung cancer and 189,597 contemporaneous controls and used routine clinical and laboratory information to predict a future lung cancer diagnosis. Importantly, the model was designed to identify disease 9–12 months before clinical diagnosis, rather than simply distinguishing patients with established cancer from healthy controls. This temporal design makes the study particularly relevant to early detection and illustrates the importance of defining the prediction time point carefully when constructing a biomarker dataset [7]. Dataset composition can also substantially influence model performance. Fan et al., investigating multiple myeloma, specifically assessed model robustness using internal and external validation datasets rather than relying exclusively on the development cohort. They reported that performance varied between validation datasets, with the divergence between datasets affecting classification results. Their findings demonstrate that a model can perform strongly under cross-validation while experiencing reduced performance when applied to a different population [8].

Data Preprocessing

Raw clinical datasets rarely enter an ML algorithm without preprocessing. Laboratory variables can differ considerably in magnitude and distribution; for example, a blood-cell count, serum enzyme concentration, and demographic variable may occupy entirely different numerical ranges. Missing observations and extreme values may further complicate model development.

Missing Data

Missing laboratory measurements can occur because a test was not clinically ordered, a specimen was inadequate, or information was unavailable in the electronic record. Removing every patient with a missing value can substantially reduce sample size and may introduce selection bias if missingness is associated with disease status or clinical characteristics. Several strategies may therefore be used, including complete-case analysis, mean or median imputation, and more sophisticated statistical or model-based imputation. The appropriate method depends on the mechanism and proportion of missingness. Importantly, information from the test set should not be used when estimating imputation parameters for the training data, since doing so can introduce data leakage.

Scaling and Transformation

Normalization or standardization may be required for algorithms sensitive to feature magnitude, particularly distance-based methods and support vector machines. A commonly used transformation is the standard z-score:

The American Journal of Engineering and Technology
(ISSN – 2689-0984)
VOLUME05ISSUE12Pages:68-87
SJIF IMPACT FACTOR (2020: 5. 32)(2021: 5. 705)(2022: 6. 456) (2023: 7. 038)
OCLC– 1121105677
Publisher:The USA Journals

$$z_i = \frac{x_i - \mu}{\sigma}$$

where x_i is the observed value, *m* is the training-set mean, and *sigma* is the training-set standard deviation. Laboratory measurements can also exhibit markedly skewed distributions. Logarithmic transformation may therefore be useful for selected variables:

$$x' = \log(x + c)$$

where c is a constant used when necessary to avoid taking the logarithm of zero.

The importance of applying preprocessing consistently was illustrated in more recent methodological work, although the principle is equally applicable to earlier biomarker studies: transformation and scaling parameters should be estimated from the development data and subsequently applied unchanged to validation data. This prevents information from the evaluation population from influencing model development.

Feature Selection and Feature Engineering

Clinical laboratory datasets may contain dozens or even hundreds of candidate variables, but not all measurements contribute independently to cancer classification. Some variables may be redundant because they measure closely related biological processes, while others may contribute little predictive information. Feature selection therefore seeks to identify a smaller subset of variables that retains useful diagnostic information while reducing model complexity. Fan et al. explicitly incorporated feature selection into their multiple myeloma ML framework. Rather than simply supplying all available routine blood biomarkers to the classifier, they selected informative variables before model construction and subsequently assessed the resulting models using internal and external validation. Their work demonstrates the practical role of feature selection in routine laboratory datasets, where the objective is not merely to maximize the number of available biomarkers but to identify a reproducible predictive subset [9].

Feature-selection approaches can broadly be divided into three categories. Filter methods evaluate features independently of the final classifier. Correlation analysis, statistical tests, mutual information, and information gain are examples. Mutual information between a feature *X* and outcome *Y* can be represented as:

$$I(X;Y) = \sum_{x,y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

A higher value indicates greater statistical dependence between the feature and outcome. Wrapper methods evaluate subsets of variables according to their performance within a selected ML algorithm. Recursive feature elimination (RFE), for example, repeatedly removes less informative variables and evaluates the resulting feature subsets. Feature engineering can additionally generate clinically meaningful composite variables. For example, hematological ratios such as the neutrophil-to-lymphocyte ratio can be derived from routinely measured cell counts. Such engineered features can provide the algorithm with representations of biological relationships that may otherwise need to be learned indirectly.

Supervised Machine Learning Algorithms

Supervised classification is particularly relevant when the objective is to distinguish patients with cancer from those without cancer. Several algorithms have been evaluated using clinical and laboratory biomarkers.

Logistic regression provides an important baseline because it produces an estimated probability of cancer:

The American Journal of Engineering and Technology
(ISSN – 2689-0984)
VOLUME05ISSUE12Pages:68-87
SJIF IMPACT FACTOR (2020: 5. 32)(2021: 5. 705)(2022: 6. 456) (2023: 7. 038)
OCLC– 1121105677
Publisher:The USA Journals

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}}$$

Its coefficients can also be interpreted in relation to the direction and magnitude of associations, making it attractive when clinical interpretability is important. Decision trees classify observations through a sequence of feature-based decisions. They can represent nonlinear relationships and produce readily understandable decision rules, although individual trees can be unstable and prone to overfitting. Random Forest addresses some of these limitations by constructing multiple decision trees using different bootstrap samples and subsets of features. Wu et al. applied Random Forest to 49 routine blood indices in 277 individuals and identified a combination of 19 indices associated with lung cancer. Ten-fold cross-validation produced sensitivity of 96.3%, specificity of 94.97%, and overall accuracy of 95.7%. The study provides a direct example of how an ensemble tree method can extract a multivariable diagnostic pattern from routine laboratory measurements [3].

Gradient boosting methods construct an ensemble sequentially, with each new learner attempting to improve the errors of the preceding ensemble. Their application to laboratory biomarker data has produced encouraging results. In a study by Wang et al. investigating early multiple myeloma diagnosis, routine laboratory variables including hemoglobin, creatinine, calcium, immunoglobulins, albumin, total protein, and the albumin-to-globulin ratio were supplied to Gradient Boosting Decision Tree, SVM, deep neural network, and Random Forest models. The Gradient Boosting Decision Tree produced the highest reported performance, with precision of 92.9%, recall of 90.0%, F1-score of 0.915, and an AUROC of 0.975 in their evaluation [9].

Ensemble and Hybrid Approaches

Ensemble learning is particularly suitable for biomarker analysis because different algorithms may capture different aspects of the underlying feature relationships. Random Forest combines multiple decision trees through aggregation, while boosting methods construct sequentially improved weak learners. Such approaches can accommodate nonlinear interactions between clinical and laboratory variables without requiring the investigator to specify every interaction explicitly. Fan et al. compared multiple ML approaches for multiple myeloma and reported that an AdaBoost-DecisionTable model produced an accuracy of 94.75%, sensitivity of 87.70%, PPV of 92.50%, F-measure of 90.00%, and AUC of 97.50% under 10-fold cross-validation. However, performance declined in some validation datasets, emphasizing that ensemble complexity does not eliminate the fundamental problem of population divergence [4].

Accordingly, algorithm selection should be considered alongside dataset quality, feature selection, interpretability, and validation rather than on predictive performance alone.

The American Journal of Engineering and Technology
(ISSN – 2689-0984)
VOLUME05ISSUE12Pages:68-87
SJIF IMPACT FACTOR (2020: 5. 32)(2021: 5. 705)(2022: 6. 456) (2023: 7. 038)
OCLC– 1121105677
Publisher:The USA Journals

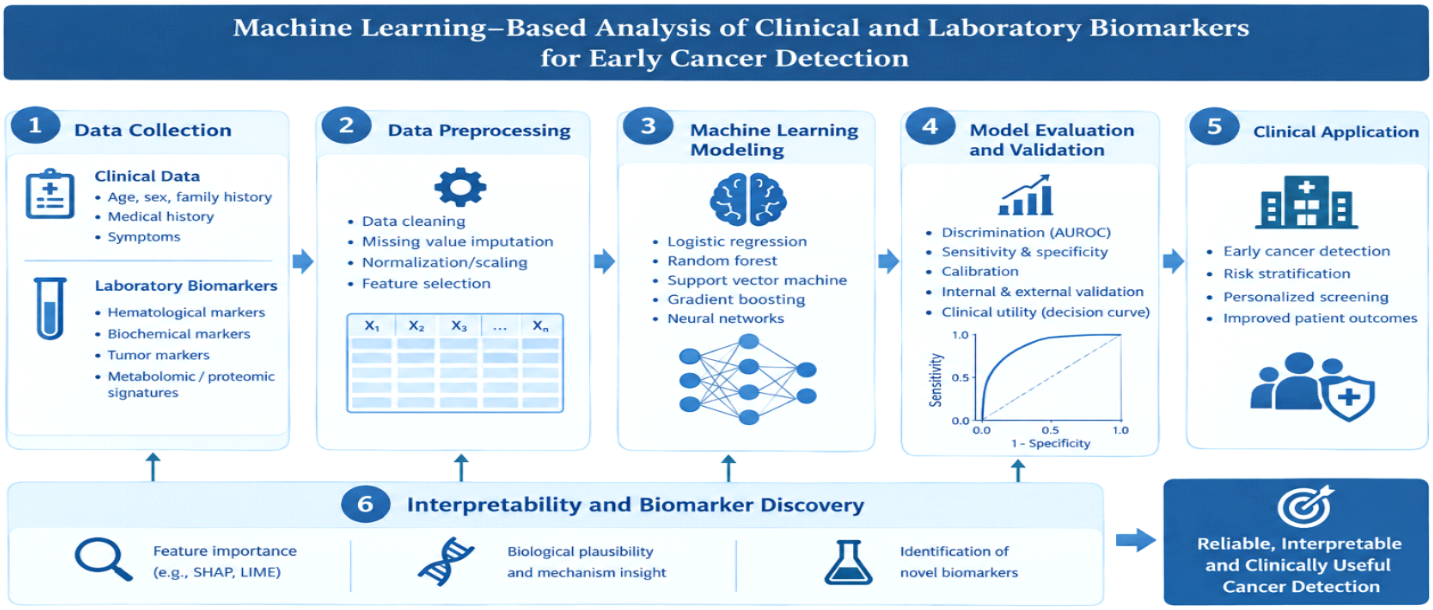


Figure 1. Integrated framework for machine learning-based analysis of clinical and laboratory biomarkers in early cancer detection.

Figure 1. Integrated framework for machine learning based analysis of clinical and laboratory biomarkers in early cancer detection

Cancer-Specific Applications of Machine Learning-Based Biomarker Analysis

The application of machine learning to clinical and laboratory biomarkers has been investigated across several cancer types, including colorectal, hepatocellular, ovarian, prostate, and gastric cancers. These studies differ substantially in their patient populations, laboratory variables, analytical methods, and validation strategies. Nevertheless, a common objective is to determine whether combinations of routinely available or disease-associated biomarkers can provide more informative cancer classification than individual laboratory measurements. The following studies illustrate the range of approaches used to integrate biochemical, hematological, metabolic, and tumor-associated information.

Colorectal Cancer

Colorectal cancer is particularly suitable for laboratory-based predictive modeling because several physiological changes associated with the disease can be reflected in routine blood investigations. In addition to anemia-related parameters, biochemical and lipid measurements may provide complementary information. Li et al. investigated this approach using electronic medical records from 1,164 individuals, comprising 582 patients with colorectal cancer and 582 healthy controls. Their dataset included liver enzymes, lipid profiles, complete blood counts, and tumor biomarkers. Five machine-learning algorithms were compared: logistic regression, Random Forest, k-nearest neighbors, support vector machine, and naïve Bayes. The logistic regression model produced an AUC of 0.865, with sensitivity of 89.5% and specificity of 83.5%.

The American Journal of Engineering and Technology
(ISSN – 2689-0984)
VOLUME05ISSUE12Pages:68-87
SJIF IMPACT FACTOR (2020: 5. 32)(2021: 5. 705)(2022: 6. 456) (2023: 7. 038)
OCLC– 1121105677
Publisher:The USA Journals

Interestingly, the four features receiving the largest model weights were CEA, hemoglobin, lipoprotein(a), and high-density lipoprotein [10]. The investigators subsequently developed a reduced model using these four variables. Although its AUC of 0.849 was slightly lower than the 12-variable logistic regression model, it retained substantial discriminatory capability. The reduced model performed particularly well in identifying late-stage colon cancer, with an AUC of 0.905, but its performance was lower for early colon cancer, where the AUC was 0.801.

This difference is important for the interpretation of ML-based cancer detection. A model that distinguishes established or advanced disease very effectively does not necessarily provide equivalent performance for early-stage detection. Consequently, cancer stage should be considered explicitly when evaluating laboratory-based predictive models. The study also illustrates the potential advantage of combining cancer-associated markers with routine laboratory variables. CEA supplied disease-specific information, whereas hemoglobin and lipid-related variables contributed complementary information. The findings therefore support a multimarker approach in which laboratory measurements are not necessarily required to be individually specific for malignancy to contribute to a useful predictive model.

Hepatocellular Carcinoma

Hepatocellular carcinoma provides a useful example of the limitations of relying on a single tumor marker. AFP, DCP, and AFP-L3 are established serum biomarkers used in HCC assessment, but their diagnostic performance can vary considerably depending on tumor characteristics and patient population. Sato et al. developed a machine-learning framework using real-world clinical data from 539 patients with HCC and 1,043 individuals without HCC. Their approach automatically optimized classifiers and hyperparameters using grid search. Among the evaluated approaches, gradient boosting achieved the highest predictive accuracy, reaching 87.34%, with an AUC of 0.940. The investigators also directly compared the ML model with conventional tumor biomarkers. Using established cutoffs, AFP achieved an accuracy of 70.67% and AUC of 0.766, while DCP achieved an accuracy of 74.91% and AUC of 0.644. AFP-L3 produced an accuracy of 71.05% and AUC of 0.683. Thus, the multivariable ML model reduced the misclassification rate relative to individual tumor markers [11]. The significance of this study extends beyond its numerical performance. The investigators explicitly framed HCC prediction as a multifactorial problem, in which cancer status cannot necessarily be represented adequately by one biomarker.

Ovarian Cancer

Ovarian cancer presents a different biomarker challenge because conventional serum markers, particularly CA-125, can be elevated in benign gynecological conditions and may have limited sensitivity in early disease. ML approaches have therefore been investigated to integrate several serum and clinical variables. Zhu et al. developed an ML-based approach using 22 conventional laboratory indicators and three demographic characteristics. The investigators analyzed retrospective samples and compared several algorithms, including logistic regression, Fisher's linear discriminant analysis, support vector machine, Random Forest, and artificial neural networks. Importantly, the study included an independently recruited prospective validation cohort rather than evaluating the models solely within the development dataset. The Random Forest model achieved an AUC of 0.989 in the prospective validation cohort, while SVM achieved an AUC of 0.985 [12]. The investigators also examined whether the models could distinguish ovarian cancer from conditions that would create a more clinically realistic diagnostic problem than comparison with healthy controls alone. The analysis included benign ovarian disease and other gynecological malignancies, thereby addressing potential diagnostic

The American Journal of Engineering and Technology
(ISSN – 2689-0984)
VOLUME05ISSUE12Pages:68-87
SJIF IMPACT FACTOR (2020: 5. 32)(2021: 5. 705)(2022: 6. 456) (2023: 7. 038)
OCLC– 1121105677
Publisher:The USA Journals

confounding. This is important because a model trained only against healthy individuals may substantially overestimate its practical diagnostic performance. Another distinctive approach was reported by Li et al., who investigated ovarian cancer using serum spectral fingerprints generated by an array of functionalized carbon nanotube sensors. The study included 269 serum samples, and several ML classifiers were trained to discriminate high-grade serous ovarian carcinoma from other diseases and healthy individuals. The best-performing classifier achieved 87% sensitivity at 98% specificity, compared with 84% sensitivity at 98% specificity for the clinical screening approach based on CA-125 and transvaginal ultrasonography [13].

Prostate Cancer

Prostate cancer provides a particularly useful example of ML-based biomarker analysis because PSA is already widely established but has limited specificity. The challenge is therefore not necessarily to replace PSA but to determine whether additional clinical information and PSA trajectories can improve discrimination. Nitta et al. investigated 512 patients who underwent prostate biopsy after PSA screening. Their ML models used age, PSA concentration, prostate volume, white blood cell count in urine, and serial changes in PSA over two years. Artificial neural networks, Random Forest, and SVM were compared with PSA concentration, PSA density, and PSA velocity [14]. When two years of serial PSA measurements were used, the ANN, Random Forest, and SVM models produced AUCs of **0.69, 0.64, and 0.63**, respectively. These values exceeded the AUCs for PSA concentration (0.53), PSA density (0.41), and PSA velocity (0.55). The corresponding accuracies of the three ML models ranged from 71.6% to 72.1%, compared with 39.1% for PSA alone. An important observation was that adding a third annual PSA measurement produced only modest changes in performance. This finding suggests that simply increasing the number of longitudinal measurements does not guarantee better prediction; the analytical representation of biomarker trajectories is also important. A separate study by Liu et al. developed a neural-network model using PSA, free PSA, age, and the free-PSA-to-total-PSA ratio [15].

Gastric Cancer

Gastric cancer has increasingly been investigated using metabolomic approaches because metabolic alterations associated with malignant transformation can produce complex patterns in blood. These patterns are difficult to characterize using conventional single-marker analysis but can be suitable for ML-based classification. A large multicenter investigation by Xu et al. examined plasma metabolic fingerprints for gastric cancer diagnosis. The retrospective component included 1,944 participants from seven centers, while an additional 264 participants were included in a prospective cohort. Plasma metabolic fingerprints were generated using nanoparticle-enhanced laser desorption/ionization mass spectrometry, and several ML methods were evaluated, including neural networks, ridge regression, LASSO regression, SVM, and Random Forest. In the retrospective discovery and independent verification datasets, the ML models produced AUC values ranging from 0.862 to 0.988. The investigators subsequently constructed a 21-metabolite panel. Its AUC ranged from 0.921 to 0.971 in the discovery dataset and from 0.907 to 0.940 in the independent verification dataset. Importantly, the panel was also evaluated prospectively, where the diagnostic AUCs ranged from 0.855 to 0.918 using the original mass-spectrometry platform and from 0.856 to 0.916 using ultra-performance liquid chromatography-mass spectrometry [16]. The multicenter and prospective components are particularly relevant. A model that performs well in one laboratory may be affected by instrumentation, population characteristics, sample handling, or analytical procedures when transferred elsewhere. The use of multiple centers and a prospective cohort therefore provides stronger evidence of reproducibility

The American Journal of Engineering and Technology
(ISSN – 2689-0984)
VOLUME05ISSUE12Pages:68-87
SJIF IMPACT FACTOR (2020: 5. 32)(2021: 5. 705)(2022: 6. 456) (2023: 7. 038)
OCLC– 1121105677
Publisher:The USA Journals

than a single-center retrospective analysis. This study also illustrates the distinction between routine laboratory biomarkers and high-dimensional laboratory signatures. Conventional blood tests generally contain a limited number of predefined measurements, whereas metabolic fingerprinting produces a much larger feature space. ML becomes valuable in this setting because it can identify a smaller combination of informative metabolites from thousands of measured signals.

Evaluation and Validation of Machine Learning-Based Biomarker Models

The diagnostic performance of a machine learning model cannot be established solely from its ability to separate cancer from non-cancer samples within a development dataset. Biomarker-based models are particularly susceptible to overfitting because laboratory variables may be correlated with age, comorbidities, treatment status, and characteristics of the study population. Consequently, evaluation should consider discrimination, sensitivity, specificity, calibration, clinical utility, and, critically, performance in data that were not used during model development. External validation is especially important when models are intended for screening populations that differ from the population in which they were trained.

Discrimination and Threshold-Dependent Performance

The area under the receiver operating characteristic curve (AUROC) is widely used to evaluate the discriminatory ability of cancer detection models. For a predicted probability p_i and observed outcome y_i, the ROC curve describes the relationship between sensitivity and *1–specificity* across possible decision thresholds. The corresponding sensitivity and specificity are defined as:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

where TP, TN, FP, and FN represent true-positive, true-negative, false-positive, and false-negative classifications, respectively.

Although AUROC provides a threshold-independent summary, it does not by itself indicate whether a model is suitable for population screening. Screening applications generally require careful consideration of false-positive results because a large number of healthy individuals may be assessed. The choice of threshold therefore represents a clinical decision rather than merely a mathematical optimization problem. A validation study of a laboratory-based colorectal cancer prediction algorithm illustrates this distinction. The model used complete blood count and demographic information and was evaluated in a large, community-based population rather than only in the cohort from which the algorithm had originally been developed. At a threshold identifying approximately the highest-risk 3% of individuals, the algorithm detected 35.4% of colorectal cancers diagnosed within the following six months, with an AUROC of 0.78. Importantly, the investigators also observed that performance decreased as the interval between the laboratory measurement and subsequent cancer diagnosis increased. This temporal relationship demonstrates why a model's performance must be interpreted according to the intended prediction horizon rather than reported as a single overall accuracy value [17].

Internal Validation and Prevention of Overfitting

Internal validation estimates how consistently a model performs within the population from which the development data were obtained. Common approaches include k-fold cross-validation, repeated cross-validation, bootstrap resampling, and the use of an independent hold-out test set. In k-fold cross-validation, the dataset is divided into k subsets, with each subset serving once as the validation data while the remaining subsets are used for model training. However, cross-

The American Journal of Engineering and Technology
(ISSN – 2689-0984)
VOLUME05ISSUE12Pages:68-87
SJIF IMPACT FACTOR (2020: 5. 32)(2021: 5. 705)(2022: 6. 456) (2023: 7. 038)
OCLC– 1121105677
Publisher:The USA Journals

validation does not eliminate all sources of bias. Feature selection, normalization, imputation, and hyperparameter optimization must be performed within the training portion of each fold. If these procedures are performed before the data are divided, information from the validation observations can indirectly influence the model, producing an overly optimistic estimate of performance.

The problem becomes particularly important in cancer biomarker studies because datasets frequently contain substantially more non-cancer than cancer observations. Wang et al. developed machine-learning models using tumor-marker measurements, age, and sex in 27,938 asymptomatic individuals, of whom 342 were subsequently diagnosed with cancer. Because the real-world dataset had an approximately 1:100 cancer-to-non-cancer ratio, the investigators used repeated random subsampling and cross-validation rather than relying on the original class distribution during model development [18].

External and Temporal Validation

External validation provides a stronger assessment of generalizability because the model is tested on patients who were not involved in model construction and who may differ in demographic characteristics, disease prevalence, laboratory procedures, or clinical practice. Ideally, validation should involve a separate institution or geographic population, while temporal validation can assess whether performance remains stable when the model is applied to patients treated or tested at a later period. The importance of independent validation is particularly evident in blood-based molecular cancer detection. Chen et al. evaluated a machine-learning classifier based on circulating tumor DNA methylation using samples from the Taizhou Longitudinal Study. The broader cohort contained more than 123,000 initially healthy individuals who were followed for subsequent cancer occurrence; the reported analysis included 605 asymptomatic individuals, of whom 191 subsequently developed one of five cancers. The classifier was trained using methylation-derived features and evaluated for its ability to identify cancer before conventional diagnosis. Although high detection rates were reported, the investigators emphasized the need for further longitudinal studies to confirm the findings [19]. This distinction is important for biomarker-based machine learning: identifying molecular differences in samples collected shortly before diagnosis is not equivalent to demonstrating effective population screening years before disease becomes clinically apparent. The temporal separation between biomarker measurement and eventual diagnosis therefore represents an important component of validation.

Calibration and Clinical Utility

Discrimination describes whether patients with cancer tend to receive higher predicted risks than those without cancer, whereas calibration evaluates whether predicted probabilities correspond to observed frequencies. A model that assigns a predicted cancer probability of 0.20 should, ideally, identify a group in which approximately 20% of individuals actually have the outcome when assessed over the relevant prediction period. Calibration can be assessed using calibration plots, calibration intercepts and slopes, and related statistical measures. Poor calibration may occur even when AUROC remains high, particularly when a model is transported to a population with a different disease prevalence.

Clinical utility provides another level of evaluation. Decision curve analysis (DCA) estimates the net benefit of using a model across clinically relevant probability thresholds. Conceptually, the net benefit can be represented as

$$NB = \frac{TP}{N} - \frac{FP}{N}\left(\frac{p_t}{1-p_t}\right),$$

The American Journal of Engineering and Technology
(ISSN – 2689-0984)
VOLUME05ISSUE12Pages:68-87
SJIF IMPACT FACTOR (2020: 5. 32)(2021: 5. 705)(2022: 6. 456) (2023: 7. 038)
OCLC– 1121105677
Publisher:The USA Journals

where N is the total number of individuals and p_t is the selected probability threshold. This approach incorporates the relative consequences of false-positive and false-negative classifications rather than treating all errors as equivalent. For cancer screening, this is particularly relevant because even a modest false-positive rate can generate substantial numbers of unnecessary investigations when millions of asymptomatic individuals are screened. Consequently, an apparently high AUROC should not be interpreted as evidence of clinical usefulness without considering calibration, threshold selection, disease prevalence, and downstream diagnostic consequences.

Interpretability and Clinical Relevance of Biomarker-Based Machine Learning

The clinical value of a machine-learning model depends not only on predictive performance but also on whether its predictions can be understood and appropriately interpreted by healthcare professionals. This is particularly important for biomarker-based cancer detection because laboratory variables have established physiological meanings. A model that identifies a laboratory parameter as predictive should therefore provide some indication of how that parameter contributes to the prediction and whether its contribution is consistent with known biological or clinical relationships. Interpretability can also help identify spurious associations, improve clinician confidence, and support the subsequent investigation of potentially useful biomarkers. Many commonly used machine-learning algorithms, particularly ensemble and neural-network models, can represent nonlinear relationships and interactions that are difficult to interpret directly. Feature-importance measures provide one approach to addressing this problem, but conventional importance scores generally indicate how much a variable contributes to model performance without explaining the direction or patient-specific effect of that variable. Shapley Additive Explanations (SHAP) provide a more detailed framework by assigning each feature a contribution to an individual prediction. A positive SHAP value indicates that the feature moves the prediction toward the target class, whereas a negative value moves it away from that class. Consequently, SHAP can provide both global explanations, describing which variables are important across a population, and local explanations, showing why a particular patient's prediction was high or low. This distinction is important in cancer biomarker research. A laboratory parameter may have relatively high global importance while exerting different effects in different patients because of nonlinear relationships or interactions with other variables. Therefore, interpreting a biomarker solely from its overall feature ranking can conceal clinically relevant heterogeneity.

A large-scale investigation by Jia et al. demonstrated the potential of explainable machine learning for extracting clinically meaningful information from routine laboratory testing. The investigators developed XGBoost-based diagnostic models for ten cancer types using 66 laboratory parameters. The study included 737,503 individuals in the training cohort, 184,012 in a validation cohort, and a separate prospective testing cohort containing 174,894 individuals. Rather than relying exclusively on conventional tumor markers, the models incorporated hematological, biochemical, coagulation, immunological, urine, and stool measurements. SHAP analysis showed that non-tumor laboratory variables contributed substantially to several cancer models. For example, albumin, prealbumin, globulin, and fecal occult blood measurements appeared among highly weighted variables for different cancer types alongside conventional markers such as AFP, CEA, PSA, and CA19-9 [20]. For some cancers, particularly gastric, biliary, thyroid, and urological malignancies, non-tumor markers occupied prominent positions among the most influential features. These findings illustrate an important advantage of machine-

The American Journal of Engineering and Technology
(ISSN – 2689-0984)
VOLUME05ISSUE12Pages:68-87
SJIF IMPACT FACTOR (2020: 5. 32)(2021: 5. 705)(2022: 6. 456) (2023: 7. 038)
OCLC– 1121105677
Publisher:The USA Journals

learning approaches: clinically useful predictive information may be distributed across multiple routine measurements rather than concentrated in a single tumor-specific biomarker. At the same time, feature importance should not automatically be interpreted as evidence that a biomarker causes cancer or represents a disease-specific biological mechanism. A variable can be predictive because it reflects inflammation, organ dysfunction, nutritional status, or another process associated with the disease.

Interpretability is particularly valuable when a model produces an individualized risk estimate. Consider two patients with identical predicted cancer probabilities but different laboratory profiles. A global feature-ranking analysis might indicate that albumin, platelet count, and a tumor marker are among the most important variables. A local explanation can instead determine which of these variables actually contributed to the prediction for each individual. This distinction has practical implications for clinical decision support. An interpretable model could present a predicted risk together with the major factors increasing or decreasing that prediction, allowing the clinician to assess whether the output is consistent with the patient's broader clinical context. Such explanations may also identify predictions that warrant caution, for example when a high-risk classification is driven primarily by a nonspecific laboratory abnormality. Other explanation methods can complement SHAP [21]. Local Interpretable Model-Agnostic Explanations (LIME), for example, approximate a complex model locally using a simpler interpretable model. Jansen et al. demonstrated the application of both LIME and SHAP to machine-learning predictions in breast cancer survival analysis and found generally consistent explanations of feature contributions.

For practical deployment, interpretability should be integrated with rather than added after model development. A clinically useful system should ideally provide a calibrated risk estimate, relevant contributing biomarkers, an explanation of the direction of their contributions, and information about the population in which the model was validated. Such a system can function as a decision-support tool rather than as an autonomous diagnostic replacement. Interpretability is also important for identifying model failure modes. If a model consistently assigns high cancer risk because of a variable associated with a particular hospital, laboratory analyzer, or patient subgroup, explanation techniques may reveal this dependence. This can prompt investigators to examine whether the model has learned a genuine disease-related pattern or a characteristic of the dataset. Ultimately, the objective of explainable machine learning in biomarker-based cancer detection is not simply to make a sophisticated algorithm easier to visualize. Its greater value lies in connecting prediction, biomarker contribution, biological plausibility, and clinical reasoning. When combined with rigorous external validation, interpretability can help determine whether a model has learned clinically meaningful patterns and whether its predictions can be responsibly incorporated into early cancer detection workflows.

Conclusion

Machine learning provides a valuable framework for integrating heterogeneous clinical and laboratory biomarkers for cancer detection, particularly when disease-associated information is distributed across multiple routine measurements rather than represented by a single biomarker. Clinical characteristics, hematological indices, biochemical parameters, inflammatory markers, metabolic profiles, and conventional tumor-associated biomarkers can be jointly analyzed to identify patterns that may not be apparent through conventional single-marker approaches. Evidence from colorectal, hepatic, ovarian, prostate, gastric, and other cancer settings demonstrates the potential of such models to improve discrimination between cancer and non-cancer populations and to

The American Journal of Engineering and Technology
(ISSN – 2689-0984)
VOLUME05ISSUE12Pages:68-87
SJIF IMPACT FACTOR (2020: 5. 32)(2021: 5. 705)(2022: 6. 456) (2023: 7. 038)
OCLC– 1121105677
Publisher:The USA Journals

support earlier identification of individuals requiring further diagnostic assessment. However, predictive performance alone is insufficient to establish clinical usefulness. Appropriate preprocessing, feature selection, prevention of data leakage, robust internal validation, and independent external validation are essential for determining whether a model has learned generalizable disease-related patterns rather than characteristics specific to its development cohort. Evaluation should extend beyond AUROC and accuracy to include sensitivity, specificity, calibration, stage-specific performance, temporal prediction horizons, and clinical utility. These considerations are particularly important for early detection, where the distinction between identifying established disease and detecting cancer at an earlier, clinically actionable stage is substantial. Interpretability represents another important component of clinically relevant machine learning. Techniques such as feature-importance analysis, SHAP, and other model-explanation approaches can identify the variables contributing most strongly to predictions and can facilitate examination of their biological and clinical plausibility. Nevertheless, predictive importance should not be interpreted as proof of causation or as definitive evidence that an individual variable constitutes an independent cancer biomarker. Instead, machine learning can serve as a mechanism for generating and prioritizing biomarker hypotheses for subsequent biological and clinical investigation. Overall, the integration of machine learning with routinely available clinical and laboratory data represents a promising approach for developing scalable cancer risk-assessment and early-detection strategies. Future progress will depend on larger and more diverse prospective cohorts, standardized laboratory measurements, rigorous external validation, transparent reporting, and evaluation within real clinical workflows. The most clinically meaningful systems will therefore be those that combine predictive accuracy with robust validation, interpretability, biological plausibility, and demonstrable clinical utility.

REFERENCES

1. World Health Organization. (2022). A short guide to cancer screening: increase effectiveness, maximize benefits and minimize harm. World Health Organization. Regional Office for Europe.
2. Hornbrook, M. C., Goshen, R., Choman, E., O'Keeffe-Rosetti, M., Kinar, Y., Liles, E. G., & Rust, K. C. (2017). Early colorectal cancer detected by machine learning model using gender, age, and complete blood count data. Digestive diseases and sciences, 62(10), 2719-2727.
3. Wu, J., Zan, X., Gao, L., Zhao, J., Fan, J., Shi, H., & Xie, X. (2019). A machine learning method for identifying lung cancer based on routine blood indices: qualitative feasibility study. JMIR medical informatics, 7 (3), e13476.
4. Fan, G., Cui, R., Zhang, R., Zhang, S., Guo, R., Zhai, Y., ... & Wang, Q. (2022). Routine blood biomarkers for the detection of multiple myeloma using machine learning. International Journal of Laboratory Hematology, 44(3), 558-566.
5. Ahamad, M. M., Aktar, S., Uddin, M. J., Rahman, T., Alyami, S. A., Al-Ashhab, S., ... & Moni, M. A. (2022). Early-stage detection of ovarian cancer based on clinical data using machine learning approaches. Journal of personalized medicine, 12(8), 1211.
6. Ahamad, M. M., Aktar, S., Uddin, M. J., Rahman, T., Alyami, S. A., Al-Ashhab, S., ... & Moni, M. A. (2022). Early-stage detection of ovarian cancer based on clinical data using machine learning approaches. Journal of personalized medicine, 12(8), 1211.
7. Gould, M. K., Huang, B. Z., Tammemagi, M. C., Kinar, Y., & Shiff, R. (2021). Machine learning for early lung cancer identification using routine clinical and

The American Journal of Engineering and Technology
(ISSN – 2689-0984)
VOLUME05ISSUE12Pages:68-87
SJIF IMPACT FACTOR (2020: 5. 32)(2021: 5. 705)(2022: 6. 456) (2023: 7. 038)
OCLC– 1121105677
Publisher:The USA Journals

laboratory data. American journal of respiratory and critical care medicine, 204(4), 445-453.

8. Fan, G., Cui, R., Zhang, R., Zhang, S., Guo, R., Zhai, Y., ... & Wang, Q. (2022). Routine blood biomarkers for the detection of multiple myeloma using machine learning. International Journal of Laboratory Hematology, 44(3), 558-566.
9. Yan, W., Shi, H., He, T., Chen, J., Wang, C., & Liao, A. (2021). Employment of artificial intelligence based on routine laboratory results for the early diagnosis of multiple myeloma. Front Oncol. 2021; 11: 608191.
10. Li, H., Lin, J., Xiao, Y., Zheng, W., Zhao, L., Yang, X., ... & Liu, H. (2021). Colorectal cancer detected by machine learning models using conventional laboratory test data. Technology in Cancer Research & Treatment, 20, 15330338211058352.
11. Sato, M., Morimoto, K., Kajihara, S., Tateishi, R., Shiina, S., Koike, K., & Yatomi, Y. (2019). Machine-learning approach for the development of a novel predictive model for the diagnosis of hepatocellular carcinoma. Scientific reports, 9(1), 7704.
12. Ahamad, M. M., Aktar, S., Uddin, M. J., Rahman, T., Alyami, S. A., Al-Ashhab, S., ... & Moni, M. A. (2022). Early-stage detection of ovarian cancer based on clinical data using machine learning approaches. Journal of personalized medicine, 12(8), 1211.
13. Kim, M., Chen, C., Wang, P., Mulvey, J. J., Yang, Y., Wun, C., ... & Heller, D. A. (2022). Detection of ovarian cancer via the spectral fingerprinting of quantum-defect-modified carbon nanotubes in serum by machine learning. Nature Biomedical Engineering, 6(3), 267-275.
14. Nitta, S., Tsutsumi, M., Sakka, S., Endo, T., Hashimoto, K., Hasegawa, M., ... & Nishiyama, H. (2019). Machine learning methods can more efficiently predict prostate cancer compared with prostate-specific antigen density and prostate-specific antigen velocity. Prostate international, 7(3), 114-118.
15. Perera, M., Mirchandani, R., Papa, N., Breemer, G., Effeindzourou, A., Smith, L., ... & Smith, E. (2021). PSA-based machine learning model improves prostate cancer risk stratification in a screening population. World journal of urology, 39(6), 1897-1902.
16. Xu, Z., Huang, Y., Hu, C., Du, L., Du, Y. A., Zhang, Y., ... & Cheng, X. (2023). Efficient plasma metabolic fingerprinting as a novel tool for diagnosis and prognosis of gastric cancer: a large-scale, multicentre study. Gut, 72(11), 2051-2067.
17. Schneider, J. L., Layefsky, E., Udaltsova, N., Levin, T. R., & Corley, D. A. (2020). Validation of an algorithm to identify patients at risk for colorectal cancer based on laboratory test and demographic data in diverse, community-based population. Clinical Gastroenterology and Hepatology, 18(12), 2734-2741.
18. Wang, H. Y., Chen, C. H., Shi, S., Chung, C. R., Wen, Y. H., & Wu, M. H. Improving multi-tumor biomarker health check-up tests with machine learning algorithms. Cancers. 2020.
19. Chen, X., Gole, J., Gore, A., He, Q., Lu, M., Min, J., ... & Jin, L. (2020). Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. Nature communications, 11(1), 3475.
20. Jia, Y., Liu, Z., Guo, J., He, C., Zhou, X., Xue, M., ... & Liu, S. (2023). Machine Learning and Bioinformatics Analysis for Laboratory Data in Pan-Cancers Detection. Advanced Intelligent Systems, 5(12), 2300283.
21. Jansen, T., Geleijnse, G., Van Maaren, M., Hendriks, M. P., Ten Teije, A., & Moncada-Torres, A. (2020). Machine learning explainability in breast cancer survival. In Digital Personalized Health and Medicine: Proceedings of MIE 2020 (pp. 307-311). 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP: SAGE Publications.