Modeling of Non-Specific Liver Biomarkers for the Early Screening of Ovarian Cancer

Prasenjit Kundu, Sayani Ghosh, Devjyoti Das, Arya Bhattacharya, and Anupam Bhattacharya*

Institute of Engineering and Management, University of Engineering and Management, Kolkata.

Abstract. Ovarian Cancer (OC) cases are exponentially growing across the world and the mortality rate are also very alarming. The present screening and detection methods of OC cases includes a series of physical tests, imaging test followed by biopsy. Such tests are time consuming, costly and painful. This contributory study attempt to identify and modeling of various non-specific Liver bio-markers and its subsequent impacts on the early detection of the OC risk factors. Blood composed of various biomolecules such as plasma, proteins, osmotic pressure, hormones, enzymes, and antibodies. A secondary blood dataset was preprocessed and analyzed in this study followed by extraction of the key significant predictors for the modeling of various non-specific blood markers for the early screening of OC cases. Albumin, Alkaline Phosphate, Direct Bilirubin and Globulin were found as significant predictors biomolecules from the dataset for OC cases detection using rankers and best first attribute selection method. The study subsequently framed three machine learning models named Instance- Based k-nearest neighbors (IBK), Random Forest (RF) and Logistic Regression (LR) classifiers with each 10-fold cross validation and compares the model performances for the selection of the best framework for the early detection of OC cases. The study evident that the accuracy of the IBK, Random Forest and Logistic Regression frameworks are 59.8%, 69.6% and 73.3% respectively with ROC value0.60, 0.75 and 0.77 respectively. That evident the performance of LR is best among other models for the accurately detection of the cases of OC with high true positive rate (76.4%) and low false positive rate. The study shows a pathway on combining features from various non-specific biomarkers for screening the risk factor of OC cases at early stage. The authors are hopeful such low cost, less invasive approaches with promising pattern finding methods of machine learning, will appreciate by medical professionals through clinical studies and tests that will help to detect OC cases at very beginning for better treatment outcome.

Keywords: Biomarkers, Machine Learning, Bio-molecules, Cross Validation

*Corresponding author: prasenjit.kundu@iem.edu.in

1 Introduction

Among the deadliest cancers in women, ovarian cancer usually presents quietly, without any warning signs and is typically discovered at an advanced stage. Early presenting symptoms are often absent in many women, and therefore, early presentation is challenging [1]. The current tools to screen for the disease, including ultrasound, CA-125 testing and biopsy, are of some use but have limitations. They can be costly, painful and cannot always detect the presence of the disease in early stages [2]. As a result, there is an increasing demand for inexpensive and non-invasive tools to assist in the early detection of ovarian cancer.

Blood biomarkers are currently being investigated as early predictors of illness. Because blood holds so many different types of molecules that can reveal what is going on within the body, it potentially provides valuable information about emerging health conditions [3]. Although CA-125 and HE4 are the most established markers for ovarian cancer, their paradigms of early detection are not always accurate and could be influenced by other medical conditions [4]. This has prompted researchers to investigate the potential of ordinary, non-specific serum markers for the early detection of ovarian cancer.

2 Literature Review

It has been observed that certain liver-derived proteins, such as albumin, alkaline phosphatase, bilirubin and globulin, may be altered in response to inflammation, metabolic stress or the progression of disease [5]. Even if not specific to ovarian cancer, such markers can still be informative since cancer often leads to broad changes throughout the body. Low levels of albumin or abnormal liver enzymes might be associated with cancer development [6], probably indicating poor status. This has prompted researchers to wonder whether ordinary blood tests, already commonly available, could be used as part of early cancer screening in a novel way.

Machine learning has emerged as an important tool for medical research because it can detect patterns in data that might not be apparent to researchers using traditional approaches. These methodologies have also been applied for prognosis prediction and disease subtyping with promising performance [7]. Algorithms such as logistic regression and random forests or k-nearest neighbors can examine multiple biomarkers simultaneously, using which combinations are most useful for prediction [8]. Feature selection also provides information on which are the most informative biomarkers for diagnosis [9].

Noteworthy is that the machine learning approach for ovarian cancer detection may also rely on complicated genetic data or imaging exams that are not available in every healthcare provider [10]. To date, relatively few studies have focused on using simple markers in conjunction with machine learning to aid in early ovarian cancer detection. This offers a significant research opportunity for low-cost, minimally invasive blood tests that are prevalent in practice and now have computational models supporting more informative results.

This research project is different than past research on biomarkers for ovarian cancer detection in many ways. Past studies focused on using specific/commonly used biomarkers for detecting ovarian cancer like CA-125 and HE4, which can be useful in the clinic but have limitations for early disease detection and/or are difficult to obtain in low-resource environments.

The present study investigates using non-specific liver function tests (e.g., albumin, alkaline phosphatase, globulin/direct bilirubin) that are not typically associated with the diagnosis of ovarian cancer. The innovation in this project is to repurpose these common types of laboratory tests as potential sources of predictive data and pair them with machine learning techniques to identify predictive patterns that might be hidden.

In addition, this research project implements a low-cost, non-invasive method for the screening of ovarian cancer and provides a comparison of several different types of machine learning models with a focus on the effect of non-linear algorithms like Random Forest. The combination of accessibility, innovation, and computational modeling creates a clear distinction between this research project and traditional biomarker-based research.

3 Research Gap

Although numerous studies exist investigating biomarkers in ovarian cancer, the majority of previous research focused on widely recognized markers like CA-125 and HE4. However, these markers can be limited with regard to early-stage ovarian cancer detection and performance in certain subgroups of patients. Little is known about the association between non-specific liver biomarkers (albumin, alkaline phosphatase, bilirubin and globulin) and ovarian cancer as a pre-clinical marker.

A further deficiency is the dearth of machine learning models developed for analyzing these routine enzymatic markers. Most machine-learning research relies on sophisticated genetic or imaging data that is costly and not readily available in all clinical settings. That, however, is on the condition that simple standard blood test markers can be merged by machine learning to create low-cost tools for early screening, which have not been researched enough. Furthermore, few analyses have contrasted different machine learning models to see which algorithms are the most adequate for these kinds of biomarkers. It's also not known how well these models would hold up away from the research lab, in real clinical practice.

This article aims to close this gap by identifying relevant non-specific liver biomarkers, and investigating different machine-learning models to assess their capability in enhancing the early ovarian cancer detection process.

4 Research Objective

- To investigate the potential contribution of non-specific liver biomarkers to early ovarian cancer screening in routine blood test results.
- To discover and screen the most important liver-related markers (albumin, alkaline phosphatase, globulin and direct bilirubin) by employing our systematic methods of data preprocessing and feature selection.
- To construct and compare several models, such as logistic regression, k-nearest neighbours, and random forest classifiers, applied to the risk prediction of ovarian cancer using the marker variables.
- To evaluate the possibility of low-cost, non-invasive biochemical markers and machine-learning patterning being used as a supplement for early ovarian cancer detection in real practice.

5 Research Methodology

This is a quantitative data-driven study which outlines research towards the evaluation of the predictive ability of non-specific liver parameters for early risk detection of ovarian cancer based on machine-learning methodology.

5.1 Data Source and Study Design

The study used a secondary clinical blood biomarker dataset from Kaggle [11], which consisted of liver-function-related biochemical variables and labels of the status of ovarian cancer. It was comprised of biochemistry covariates associated with liver function and a binary outcome representing ovarian cancer status (with 0 being non-cancer and 1 being disease). Secondary data were utilised, thereby avoiding the need to involve patients directly in analyses.

5.2 Data Preprocessing

The dataset was initially checked for missing values, anomalous usage, and outliers. Cleaning methods were employed to ensure the quality of the data. Numerical attributes were normalised to mitigate the effects of scale, and machine-learning algorithms then performed better. Class labels were checked for consistency across the dataset.

5.3 Feature Selection

Several feature-selection methods, such as Best-Fit selection, Greedy Stepwise search, and Ranker, were used to identify the most predictive predictors. These strategies contributed to dimension reduction and the identification of the most informative liver biomarkers. According to the uniform selection across all methods, albumin, alkaline phosphatase, globulin, and direct bilirubin were selected to develop the model. To reduce dimensionality and identify the most predictive biochemical markers, multiple feature-selection approaches, including Best-Fit, Greedy Stepwise and Ranker, were applied. The biomarkers consistently identified across these methods are presented in Table 1.

Table 1. Selected non-specific liver biomarkers identified using feature-selection methods

Feature Selection Method	Selected Biomarkers	Ranking Score (if applicable)
BestFit Method	ALB, ALP, GLO	—
Greedy Stepwise Method	ALB, ALP, GLO	—
Ranker Method	ALB ALP GLO DBIL	0.1203 0.0811 0.0501 0.0306

5.4 Model Development

Three supervised dimensionality reduction algorithms are chosen for classification: Logistic Regression, k-Nearest Neighbours (IBK), and Random Forest. Logistic Regression was selected for interpretability, as well as IBK and Random Forest, to account for non-linear dependencies between biomarkers for the early-stage Ovarian Cancer case detection.

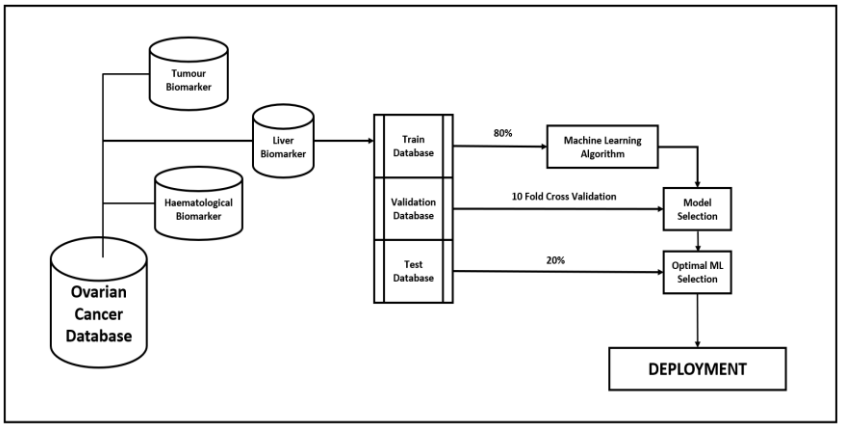

Fig 1: Block Diagram of the proposed model

5.5 Model Validation

The model was validated by two validation approaches. Stability and sampling bias were checked first by ten-fold cross-validation for the full dataset. Second, an 80:20 randomly split into train–test was conducted to evaluate real-world predictive performance. Both the validation methods guaranteed stability and overfitting prevention.

5.6 Performance Evaluation

Model performance was measured with various assessment indices, including accuracy, sensitivity, specificity, kappa statistics, root mean square error (RMSE) and receiver operating characteristic (ROC) values. Treatment for cancer and non-cancer cases was also evaluated using the confusion matrix.

5.7 Interpretation and Analysis

Comparisons between machine learning models were made to find out which model achieves the best performance. A particular emphasis on sensitivity and false-negative rates because they are of clinical relevance for early cancer diagnosis. The findings were discussed in relation to clinical implications and predictive validity.

6 Discussion, Result and Analysis

The current study evaluated the utility of generic liver biomarkers in predicting ovarian cancer risk by applying machine-learning algorithms. The study centred on four biochemical parameters (albumin, alkaline phosphatase, globulin, and direct bilirubin) related to liver function; they were determined using systematic feature-selection approaches. These biomarkers are not disease-specific but might mirror systemic inflammation and metabolic disturbances that occur during cancerogenesis.

6.1 Features selection and biomarker importance

The best predictors were identified using various feature-selection algorithms, such as Best-Fit, Greedy Stepwise, and Ranker. Albumin, alkaline phosphatase and globulin were

consistently identified as significant predictors among all approaches, suggesting that they are highly associated with the risk of ovarian cancer. Direct bilirubin was ranked low as well, but was kept in the analysis because of its antioxidant and anti-inflammatory effects as well as its relevance to cancer outcome. These results demonstrate that directly interpretable individual biomarkers have a poor diagnostic rate, but they can contribute to predictive models in a meaningful way.

6.2 Model Performance Using Cross-Validation

Three machine learning methods: Logistic Regression, IBK (k-nearest neighbours) and Random Forest were utilised to build the predictive model by 10-fold cross-validation. Due to the poor model fit, as this measure of association is assumed beyond linear reasoning, we believe that logistic regression could not capture the complex interplay among our biochemical variables in modelling predictive response. Both IBK and Random Forest have more powerful classification ability, which emphasizes the necessity of non-liner model in the analysis of biological data. The output from the Random Forest model is shown in Figure 1, obtained using the WEKA tool with 10-fold cross-validation.

6.3 Results and Analysis for Train–Test Split

To assess real-world applicability, an 80:20 train–test split was used for validation. Both IBK and Random Forest yielded high classification accuracy, with nearly all ovarian cancer and non-cancer cases correctly classified. These models had a strong fit between predicted and actual outcomes, as indicated by high kappa statistics and ROC values near perfect. The Logistic Regression, however, performed with lower accuracy and higher misclassification rates, implying only limited predictive validity based on this dataset.
The confusion matrix analysis also demonstrated that IBK and Random Forest had no false-negative cases, which is therapeutically important as misdiagnosing ovarian cancer could lead to severe consequences. LR showed a relatively large number of false-positive and false-negative cases, thus reducing its validity as an independent screening model.

6.4 Model Interpretation and Clinical Relevance

Although IBK attains very good sensitivity, its nearest-neighbour matching may lead to overfitting, especially with small datasets. The Random Forest, however, had a more balanced and resistant performance in capturing complex, non-linear interactions between biomarkers. This indicates that the interaction of albumin, alkaline phosphatase, globulin and direct bilirubin is crucial to discriminate ovarian cancer risk.
However, Logistic Regression is still useful for clinical interpretation due to providing not only the direction but also the order of influence of each biomarker. Its lower predictive accuracy suggests it is probably more appropriate as a support than a primary screening tool. The comparative performance of the classifiers is summarised in Table 2, where Random Forest and IBK significantly outperform Logistic Regression across all evaluation metrics.

Table 2: Performance comparison of models (80%–20% Train–Test Validation)

Performance Metric	Random Forest	IBK	Logistic Regression
Correctly Classified Instances (%)	98.57	98.57	67.14

Incorrectly Classified Instances (%)	1.43	1.43	32.86
Kappa Statistic	0.9714	0.9714	0.3429
Mean Absolute Error	0.1677	0.0117	0.4121
Root Mean Squared Error	0.2005	0.0684	0.4798
ROC	1.00	1.00	0.673
Confusion Matrix	a b <-- classified as	a b <-- classified as	a b <-- classified as
	34 1 | a = 0	34 1 | a = 0	22 13 | a = 0
	0 35 | b = 1	0 35 | b = 1	10 25 | b = 1

6.5 Overall Interpretation

The results are evident that non-selective liver biomarkers, when combined, can be valuable for predicting ovarian cancer risk using machine-learning methods. The better performance of Random Forest emphasises the interest in ensemble-based non-linear models for biomedical data analysis. These results indicate the potential to integrate blood test parameters into an early, inexpensive screening algorithm, which may be particularly useful in areas where sophisticated diagnostic tools are scarce.

7 Conclusion

These results demonstrate that non-specific liver-related blood biomarkers can be utilised to facilitate early risk detection of ovarian cancer using machine-learning tools. The most significant biochemical parameters were albumin (ALT), alkaline phosphatase, and direct bilirubin, suggesting that changes in systemic inflammation and liver function are related to the risk of OC. The relative performances of the ML models revealed striking differences. Moderate prediction capability (67.14%) was achieved by Logistic Regression, with an ROC of 0.673, suggesting limited capacity to capture complex biomarker interplay. In contrast, the classification accuracy of both IBK and Random Forest models was high (98.57%), which confirms excellent predictive reliability, with a kappa value of 0.97 and an ROC value of 1.00. Additional analysis found that IBK achieved perfect sensitivity (sensitivity=1), correctly identifying all ovarian cancer cases; however, it is susceptible to nearest-neighbour overfitting. Random Forest achieved the best balance and robustness by preserving high accuracy and capturing nonlinear relationships among biomarkers. The confusion matrix analysis revealed no false-negative cases for Random Forest, being clinically highly important in the early detection of cancer.

In total, the results indicate that integrating common liver biomarkers with ensemble-based machine-learning models has the potential to present a low-cost, non-invasive and highly accurate method for early ovarian cancer prediction. The results suggest that Random Forest possibly has clinical promise as a decision-support technique, and Logistic Regression is useful for interpretation. And with additional clinical validation and incorporation of age and gamma-glutamyl transferase variables, this model may contribute to improved early screening strategies, leading to better outcomes for patients.

8 Limitations and future scope

8.1 Limitations

Our study has several limitations. First, it relies on a secondary data set, so the researchers could do little about how the data were gathered or how balanced their sample was. Second, the dataset may not be reflective of all populations, in which case the conclusions would not be generalizable to other age groups, ethnicities or medical histories. Thirdly, it has been limited to a small number of non-specific liver biomarkers, and other significant biomarkers might be overlooked. Finally, the machine-learning models were evaluated only in an available dataset, rather than validating them in actual clinical conditions; this may indicate that our findings are limited to specific populations.

The model has not been validated on independent or clinical datasets, which may limit its applicability in real-world clinical settings. The near-perfect performance observed in Random Forest and IBK models may indicate potential overfitting, especially considering the moderate dataset size. Although cross-validation was applied, further validation on independent datasets is necessary to confirm model robustness.

8.2 Future scope

The research will lead to multiple opportunities for future investigation. To further substantiate the findings of this research, larger databases representing different population groups may be required to validate this finding and increase the precision of the model. Future research will also need to use both the candidate feature sets used by this study and seek other new candidate biomarkers such as oxidative markers or hormonal & metabolic indicators which may provide a higher level of predictive accuracy. The use of imaging data and/or genetic data in conjunction with biochemical markers may also provide a way to develop advanced screening tools. Finally, the completion of future clinical trials and real-world implementation studies will be essential to provide evidence of the validity of AI machine-learning models and how these models may be employed in systematic screening programs or other clinical settings.

The proposed methodology may have applications beyond this study for other types of cancers. Because the proposed methodology is based on identifying repetitive patterns in daily blood biomarkers, the methodology is also applicable to markers that indicate physiological changes caused by diseases process (i.e., inflammation, metabolic imbalance, or organ dysfunction), which occurs in every subtype of cancer. Thus, machine-learning frameworks that have previously been used for breast cancer may be replicated for identifying early indicators of risk in other types of cancers like liver, pancreatic or colorectal cancers.

However, it is important to note that the predictive biomarkers and their relative importance may vary depending on the cancer type. Hence, disease-specific datasets and validation studies would be required before generalizing this approach to other cancers. Future research can explore this direction by integrating additional biomarkers relevant to specific malignancies.

9 Clinical implications

Validated through larger datasets of different populations can enhance the model accuracy and provide further evidence to support the discovery. Future work across candidate features will also incorporate new biomarkers such as inflammatory markers as well as

metabolic/hormonal factors to improve predictability. Combining imaging data or genetic data with biochemical markers can create advanced screening tools too. Finally, clinical trials and real-world implementation studies will be necessary to validate the accuracy of the machine-learning models and assess how those models would work in practice as part of screening programs.

This methodology has potential worldwide applications, since its basis is in established, inexpensive blood indicators, (i.e., albumin, alkaline phosphatase, globulin and bilirubin) widely accessible in both developed and developing health care systems across the globe. This makes it an ideal fit for low income nations where access to more advanced diagnostic modalities such as imaging or genetics testing is restricted. However, since this study utilized a secondary database to identify these indicators; it is likely not accurately representative of all populations. In order to generalize globally it is necessary to validate this model on multiple databases of multi-sites encompassing varying demographic/ethnic/clinical populations.

The predictive markers assembled into this model can serve as an initial screen for patients being evaluated for possible diagnosis, thereby avoiding the use of imaging and CA-125 tests until subsequent to obtaining conclusive diagnostic results. Subsequently to that CA-125 has exhibited little sensitivity for accurately identifying early stages of usually asymptomatic malignancy, but the multiple routine indicators incorporated into this integrative approach allow for a potentially complimentary cost-saving method of screening these patients.

References

1. L.A. Torre, B. Trabert, C.E. DeSantis, K.D. Miller, G. Samimi, C.D. Runowicz, M.M. Gaudet, A. Jemal, R.L. Siegel, Ovarian cancer statistics, 2018, CA Cancer J. Clin. 68(4), 284–296 (2018)
2. S. Lheureux, C. Gourley, I. Vergote, A.M. Oza, Epithelial ovarian cancer, Lancet 393(10177), 1240–1253 (2019)
3. B. Zhang, J. Wang, X. Wang, G. Liu, Circulating biomarkers for early detection of ovarian cancer, J. Cell. Mol. Med. 24(16), 9515–9527 (2020)
4. M.J. Duffy, C.M. Sturgeon, D. de Souza, D. O'Gorman, CA125 in ovarian cancer: Updating a clinically relevant biomarker, Clin. Chim. Acta 524, 46–54 (2022)
5. B. Nazha, M. Mishra, R. Pentz, T.K. Owonikoko, Albumin as a prognostic biomarker in cancer: A review, J. Oncol. 2015, 1–6 (2015)
6. G. Hou, X. Zhang, W. Gong, Liver function tests and cancer risk: Insights from clinical and epidemiological studies, Cancer Med. 9(7), 2343–2352 (2020)
7. K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Comput. Struct. Biotechnol. J. 13, 8–17 (2015)
8. J.A. Cruz, D.S. Wishart, Applications of machine learning in cancer prediction and prognosis, Cancer Inform. 2, 59–77 (2006)
9. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3, 1157–1182 (2003)
10. M. Lu, D. Zeng, S. Shao, Machine learning-based early detection of ovarian cancer using clinical and biomarker data, Artif. Intell. Med. 98, 39–47 (2019)
11. R. Mittal, Ovarian cancer blood biomarker dataset, Kaggle (2021)

https://www.kaggle.com/datasets/therishabhmittal05/ovarian-cancer-blood-biomarker-dataset