

Multi-phase feature selection for detection of epithelial ovarian cancer using ensemble machine learning techniques

Suma Palani Subramanya, Suma Kuncha Venkatapathiah

Department of Electronics and Communication Engineering, Ramaiah Institute of Technology, Affiliated to Visvesvaraya Technological University, Belagavi, India

Article Info

Article history:

Received Jan 30, 2025

Revised Sep 12, 2025

Accepted Oct 18, 2025

Keywords:

Correlation coefficient

Ensemble classifier

Machine learning

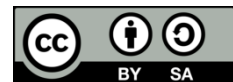
Ovarian cancer

Recursive feature elimination

ABSTRACT

Epithelial ovarian carcinoma is one of the most prevalent causes of death. Timely ovarian cancer diagnosis is significant for bettering patient outcomes and rates of survival. For prognostic and diagnostic evaluation of malignancies, AI-based machine learning algorithms are used. This novel technique is undoubtedly an effective tool that may aid in selecting the best course of action. The collection of data comprising 150 patients contained an extensive selection of clinical characteristics and markers of tumors. The recursive feature elimination (RFE) and correlation coefficient feature selection techniques were assimilated to pick the features for the machine learning model, such as age, CA-125, tumor laterality, size, tumor type, grade of tumor, and International Federation of Gynecology and Obstetrics (FIGO) stage. The study's findings indicate that the base model accuracy was around 96%, sensitivity 93%, and specificity 100%. Using ensemble classification, accuracy was around 96%, sensitivity 98%, and specificity 94% for the RFE technique. By obtaining a deeper understanding of their decision-making process, explainable artificial intelligence makes sophisticated machine learning methods easier to explain. Before beginning treatment, this research offers crucial data for the diagnosis and prognosis assessment of individuals with epithelial ovarian cancer (EOC).

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Suma Palani Subramanya

Department of Electronics and Communication Engineering, Ramaiah Institute of Technology

Affiliated to Visvesvaraya Technological University

Belagavi-590018, India

Email: sumap1994@gmail.com

1. INTRODUCTION

Ovarian cancer remains a major health concern because of the difficulties in treating it, which are mostly related to delayed identification and the intricacy of its disease detection. India is expected to see 49,644 new cases of ovarian cancer in 2025 [1]. This is a minor rise over the 43,886 instances that were predicted for 2020. Additionally, the five-year survival rate from 2015 to 2021 was 51.6% [2], [3]. These outcomes highlight the implication of early detection and therapeutic strategies, an area of continuous research need to be developed [4], [5].

The diagnostic methods that are employed to identify ovarian cancer include lower abdomen tests, medical imaging procedures such as transvaginal ultrasound (TVUS), magnetic resonance imaging (MRI), and blood tests. The most prevalent biomarkers are cancer antigen 125 (CA-125), CA19-9, carcinoembryonic antigen (CEA), and human epididymis protein 4 (HE4). The most responsible technique for determining ovarian cancer is histopathology because it examines tissue samples at the cellular level [6]. The reliability of TVUS and pelvic examination in segregating between benign and malignant tumors is very

low. Similarly, it may be difficult for MRI to reliably identify whether a tumor is malignant, but it can produce precise images of the ovaries and the surrounding tissues. HE4, carbohydrate antigen 72-4 (CA72-4), and carbohydrate antigen 125 (CA-125) are important cancer biomarkers that can identify female pelvic tumors [7]–[9].

A typical course of medication for ovarian cancer involves surgery followed by chemotherapy [10]. But most of the patients eventually experience a recurrence of the disease that is usually incurable, primarily as a consequence of the formation of drug resistance. Because of the severe side effects and expense of chemotherapy, precision medicine aims to classify patients whose malignancies are resistant to the treatment.

Artificial intelligence (AI) is essential in the therapeutic industry for cancer diagnosis and detection [11]. In a rising number of medical applications, AI is emerging as a feasible alternative for decision-making algorithms. In the medical field, machine learning is frequently used to evaluate patient data and provide early disease diagnoses. The machine receives the patient's key characteristics as input and outputs a precise diagnosis [12]. Innovative approaches using machine learning algorithms hold significant promise for diagnosing cancer and forecasting the course of disease. Many researchers have worked on machine learning algorithms in order to accurately detect ovarian cancer using these biomarkers. Arezzo *et al.* [12] used machine learning algorithms on ultrasound images to calculate the 12-month survival period for ovarian cancer patients. Further five-fold cross-validation was used to train and validate three distinct machine learning algorithms, logistic regression (LR), random forest (RF), and k-nearest neighbors (KNN) to forecast a 12-month survival period. The highest performance accuracy was 93.7%. Ziyambe *et al.* [13] applied a convolution neural network to histopathological images to predict and diagnose ovarian cancer and achieved an accuracy of 94%.

The principal intent of this research is to employ ensemble-based machine-learning algorithms to assess the pre-operative status of those diagnosed with ovarian cancer. The most important features, such as age, tumor laterality, size, tumor type, tumor grade, International Federation of Gynecology and Obstetrics (FIGO) stage, and CA-125, are selected using two feature selection techniques. There is a close association between the clinical factors and the effective tumor marker CA-125. This will allow the doctors to treat the patients appropriately, which will increase patients' longevity.

The paper is designed as follows: the prior investigations conducted for the diagnosis and detection of ovarian cancer are covered in section 2. A thorough explanation of each element of the suggested framework is given in section 3. Section 4 presents the findings and an analysis of the research. Section 5 presents the study's conclusions.

2. LITERATURE SURVEY

The recent studies employed machine learning models on significant biomarkers for the detection of ovarian cancer. Different machine learning algorithms proposed by Lavanya and Pasupathi [14] include KNN, support vector machine (SVM), decision trees (DT), followed by max voting, boosting, bagging, and stacking. Data was collected from Kaggle. To select the features minimum redundancy maximum relevance (MRMR) algorithm was used. SVM has 85% accuracy, and stacking 89%. Wibowo *et al.* [15] discussed the classification of ovarian cancer using KNN and SVM and achieved a classification accuracy of 90.47% for KNN.

Ahamad *et al.* [16] focused on ensemble models in addition to machine learning techniques to categorize between healthy and cancerous patients. Various significant Biomarkers used in the study are CA-125, HE-4, CEA, and CA19-9. Overall, this work attained an accuracy of 91%. A machine learning model, proposed by Taleb *et al.* [17], uses machine learning algorithms to progress in the precision of ovarian cancer diagnosis. The model is simulated using MATLAB 2021a. Performance is assessed using a variation of statistical metrics using the proposed model, with an accuracy of 97.16%.

Ahamad *et al.* [16] identify major blood biomarkers like CA-125, CA 19-9, CEA, and HE-4, along with other critical parameters. The study discusses the application of several machine learning methods, emphasizing the need for early identification to improve patient outcomes, including DT, RF, SVM, gradient boosting machine (GBM), LR, light gradient boosting machine (LGBM), and extreme gradient boosting (XGB), in categorizing ovarian cancer patients based on clinical data with an accuracy of 91%. Wang *et al.* [18] performed initial screening of ovarian cancer based on the risk of ovarian malignancy algorithm (ROMA). The biomarkers HE-4, CA-125 were used and obtained an AUC of 0.91 for ROMA. Bast *et al.* [19] talk about using biomarkers to diagnose ovarian cancer early, such as CA-125, microRNAs, ctDNA, and methylated DNA. A performance of 98% is achieved for all the control subjects.

Wibowo *et al.* [15] main objective was to classify ovarian cancer using machine learning procedures, namely KNN and SVM. This paper explains how well machine learning processes such as KNN and SVM categorize ovarian cancer cases; in this particular research, KNN performed better than SVM with an accuracy of 90.47%. To increase the current biomarker combination model's ability to classify ovarian cancer, Song *et al.* [20] aim to include menopausal data. The area under the ROC curve of 0.985 specifies

that the typical with menopausal data achieves better than the model without clinical data in the study's evaluation for ovarian cancer screening.

To increase the detection accuracy Abuzinadah *et al.* [21] used ensemble models such as bagging and boosting which include 50 features. Combining ensemble models lowers the variance for precise ovarian cancer detection, and the resulting accuracy of 96.87% was achieved. In order to generate a new predictive diagnosis, Paik *et al.* [22] assessed the accuracy of gradient boosting (GB) using conventional statistical methods. The research presented how machine learning techniques, particularly GB, can surpass conventional statistical methods in precisely forecasting the individual outcomes of epithelial ovarian cancer (EOC) patients with an area under the curve (AUC) value of 0.830. Nayak *et al.* [23] described RF as a machine learning classifier which resulted in an accuracy of 91%. Table 1 details about the work carried out in the recent literature.

The most recent study focused on detecting ovarian cancer solely through biomarkers. However, in order to offer a meaningful diagnostic tool for evaluating the pre-operative condition of ovarian cancers, clinical characteristics and biomarkers must be integrated. Therefore, the key contribution of this work as follows: i) to choose the best clinical characteristics that are essential for the prompt detection of ovarian cancer, two feature selection strategies are used, ii) machine learning and ensemble models are employed to achieve the early diagnosis and prognosis of ovarian cancer, and iii) machine learning algorithms are further enhanced by explainable artificial intelligence (XAI), which makes the decision-making procedures explicit, comprehensible and effective.

Table 1. Synopsis of pertinent research

Ref.	Method used	Sample size	Performance parameters	Limitations
[24]	LR, SVM, RF, DT, KNN	350 patients in the Kaggle dataset, and among the 50 parameters are age, menopause, and type.	Accuracy from RF was 90.4% with the top features selected were age, CA-125, menopause, HE-4, neutrophils (NEU), and with random features, accuracy was around 81.9%. RF gave the best accuracy and AUC, of 69.0% and 0.760, respectively.	Accurate identification of ovarian cancer necessitates a better comprehension of the disease's pathophysiology.
[25]	GB machine, SVM, RF, conditional RF (CRF), naïve Bayes (NB), neural network, and ElasticNet.	101 individuals with normal ovarian tumors and 334 patients with EOC.		Robust validation and clinical implementation initiatives, external validation in potential groups is required.
[26]	RF algorithm with a 10-fold cross-validation technique.	157 serum samples from healthy non-cancer controls and 143 from ovarian cancer patients	Five biomarkers, Apolipoprotein [Apo] A1, ApoA2, HE-4, CA-125, and cancer antigen 15-3, showed 93.71% sensitivity and 93.63% specificity	Focused solely on diagnostic accuracy rather than forecasting response to therapy or survival
[27]	An additional benefit is multivariate analysis with tumor biomarkers in addition to ultrasonography.	Women above 50 years are defined	Sensitivity and specificity of CA-125 with CA 19-9, EGFR, G-CSF, Eotaxin, IL-2R, cVCAM, and MIF were 98.2% and 98.7%, respectively.	Lack of a detailed dataset specification, missing information about the disease stages, unknown imaging inputs, or biomarkers
[28]	RF, DT, Gaussian NB, AdaBoost, and LR.	349 patients	RF classifier had the greatest validation accuracy at 99%.	Lack of clarification regarding the data sources, such as imaging, biomarkers, or clinical records. No mechanism for integrating clinical procedures
[29]	Parameter optimization and feature weighting for ovarian cancer detection. Weights maximization-least absolute shrinkage and selection operator (LASSO) regularization and adverse drug event (ADE) with cross-validation error.	235 patients	With KNN, 97.24% accuracy and 96.48% (mean) accuracy with SVM	The inability to comprehend the optimization processes makes clinical implementation difficult. Lack of testing to verify robustness on a separate cohort.
[30]	Multiclass SVM, ANN, and Naive	493 data from the cancer genome atlas program (TCGA) portal	Multiclass SVM=98%	Provides a summary of previous research. Does not assess whether the assessed machine learning models are clinically applicable.

3. METHODOLOGY

The following are the key phases that are necessary in the diagnosis of ovarian cancer: pre-processing, feature selection techniques, machine learning models, and model explanation process, as indicated in the Algorithm 1. As shown in Figure 1, pre-processing the data is the initial step in the process. It involves the dropna () method to replace the missing values of the required feature column from the dataset. Following the

pre-processing, the data is divided into two sections as testing and training. Using an assortment of feature selection and machine learning techniques, the suggested model is evaluated using standard metrics to determine the pre-operative state of ovarian cancer. To improve interpretability, the consequence of the chosen features is demonstrated using an XAI technique.

Algorithm 1. Structured machine learning workflow for ovarian cancer biomarker classification with dual feature selection strategies

1. Preprocess and load the data using
load_data ('DATASET_1.xlsx')
2. Set up features and target
X=data.drop ('CA125'); y=(data ['CA125'] >35)
3. Data split (50% testing, 50% training)
split (X, y)=X_train, X_test, y_train, y_test
4. Select features:
 - i) Based on correlation (threshold=0.1)
 - ii) Based on the recursive feature elimination (RFE)
5. Develop and assess models:
 - i) Using features chosen by correlation
 - ii) On features chosen by RFE
6. Use SHAP to interpret the optimal model.
7. Create graphs of related performance metrics

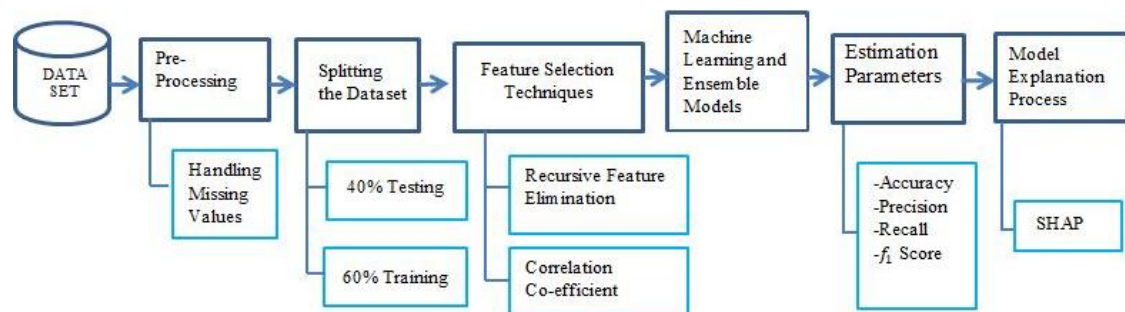


Figure 1. Flow diagram of the proposed work

3.1. Dataset

Databases from Ramaiah Medical College, Bangalore, were used in this study. It has a total of 28 characteristics, including the tumor marker CA-125. Seven most significant features are selected, such as age, tumor laterality, size, tumor type, tumor grade, and FIGO stage. There are 84 cases of ovarian cancer and 66 cases of non-ovarian cancer among the 150 individuals in the dataset.

3.2. Pre-processing

Throughout the data assessment process, we preserved as much of the original data as possible to guarantee that it could be used completely. The size column was missing 10% of the value. To deal with the missing values, the median is specifically calculated.

3.3. Separating the dataset

The dataset was collected from Ramaiah Medical College, Bangalore. It was divided into training and testing phases. Specifically, 60% of the data was used for training, while the remaining 40% was used for testing.

3.4. Feature selection techniques

The feature selection methods are a more understandable and practical way for model creation by taking pertinent information out of the raw data. The optimal feature technology can differ because of the specific dataset, the issue being solved, and the algorithms employed. It often requires a deep comprehension of the facts and extensive domain expertise. To determine the efficacy of the features that are created, iterative feature engineering necessitates testing and validation.

3.4.1. Recursive feature elimination technique

The least important characteristics are progressively removed using RFE, a backward elimination feature selection method [31]. It performs grading of each feature according to the way the model performs. By progressively removing features, RFE reduces predictor dependence.

3.4.2. Correlation coefficient

The linear relationship among features and the desired variable is measured by the correlation coefficient [31]. High correlation features around ± 1 is deemed significant. When characteristics have linear predictive power, it works best.

3.5. Machine learning and ensemble models

The features chosen by feature selection approaches are used by machine learning models, which are crucial for decision-making. Further to improve the performance from base models, ensemble approaches are used. There are two stages in this section.

3.5.1. Stage 1: machine learning models

This method allows machine learning algorithms to learn and identify patterns in the data collectively by providing them with access to the training database [32]. Various machine learning algorithms are employed in this work. These include SVM, KNN model, DT, and LR.

3.5.2. Stage 2: ensemble models

Ensemble models such as voting classifier, staking, bagging and boosting are used to progress the performance of base models [32]. When a complex dataset is unavailable, machine learning models should be used instead of deep learning models. Also, Ensemble models are found to be better performing than individual base machine learning models for the abovementioned reason. Figure 2 indicates different ensemble models obtained by combining independent base machine learning models' predictions. The ensemble models are elaborated in this section:

- Voting classifier: voting classifiers [22] are machine learning models that predict an output class based on which model has the best chance of producing the target class.
- Bagging: a meta-algorithm called bootstrap aggregating, sometimes referred to as bagging [22], aims to increase the accuracy and consistency of machine learning techniques used in analytical regression and classification. It reduces variance and helps avoid overfitting. In accordance with (1) and (2), DT techniques are typically applied with it. One special application of the model averaging approach is bagging.

$$\text{Manhattan} = \sum_{k=1}^n |p_k - q_k| \quad (1)$$

$$\text{Minkowski} = (\sum_{k=1}^n (p_k - q_k)^x)^{\frac{1}{x}} \quad (2)$$

- Stacking: in order to increase prediction performance, stacking is an ensemble learning strategy [22] in machine learning that integrates base models, also known as base learners. The goal is to create a better overall model by strategically combining the outputs of many models and utilizing their respective strengths.
- Boosting: XGB is another technique for enhancing machine learning [22]. An extreme variant of the GB technique is the extreme GB algorithm, sometimes known as XGB. GB and XGB differ primarily in that the former employs a regularization technique. It is a regularized form of the currently employed gradient-boosting technique. This explains why XGB outperforms a traditional GB method. Additionally, it performs better in datasets that contain both numerical and categorical variables.

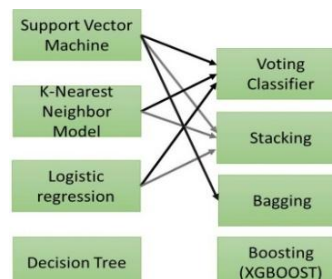


Figure 2. Block diagram of ensemble model

3.6. Performance indicators

The effectiveness of the machine learning methods that have been implemented is evaluated using four different criteria. They are:

- i) True positive (Tp): it is used to forecast the event value more precisely.
- ii) False positive (Fp): in essence, this technique is employed to ascertain the erroneous value of an occurrence.
- iii) True negative (Tn): the purpose of this metric is to predict when an event value will not occur.
- iv) False negative (Fn): this is used when the no event value is wrongly predicted.

Precision, recall, F1-score, and accuracy are the four performance indicators.

- Precision: the model's quality is referred to as precision. In simple terms, the most genuinely positive out of all favorable predictions as per (3) [23].

$$Precision = \frac{T_p}{T_p + F_p} \quad (3)$$

- Recall: the ratio can be calculated as shown in (4) [23].

$$Recall = \frac{T_p}{T_p + F_n} \quad (4)$$

- F1-score: erroneous positive and erroneous negative results are also taken into consideration in this performance. Therefore, it works effectively with both balanced and imbalanced data sets as per (5) [23].

$$F1 - score = \frac{2(Precision * Recall)}{Precision + Recall} \quad (5)$$

- Accuracy: this is computed by dividing the total number of samples by the number of examples that were correctly identified.

4. RESULTS AND DISCUSSION

The presence of ovarian cancer was determined by analyzing a number of medical characteristics that were included as part of the dataset. The most important features, such as age, CA-125, tumor laterality, size, tumor type, grade of tumor, and FIGO stage, for the analysis were selected. Table 2 gives the detailed information on the patients in the dataset.

Table 2. Key information of the patients in the dataset

Parameters	Categories	Total no. of patients (N=150)	Percentage
Age	>50	63	42
	<50	63	42
CA-125	>35 U/mL	84	56
	<35 U/mL	66	50
Size	>4.8 cm	134	89
	<4.8 cm	16	10
Tumour laterality	Unilateral	125	83
	Bilateral	25	17
Tumour type	Serous	98	65
	Endometroid	27	18
	Clear cell	6	40
	Germ cell	0	0
	Sexcord		
Grade of tumour	Brenner	4	3
	Mucinous	15	10
	Grade I	8	5
	Grade II	114	76
	Grade III	5	3
Figo stage group	Grade IV	23	15
	Stage I	60	40
	Stage II	6	4
	Stage III	82	55
	Stage IV	2	1

In this study, Python (3.8) is the core programming language. The following Python libraries are used: Pandas for loading and processing, NumPy to perform basic operations, scikit-learn for machine learning algorithms, and plots were created using Matplotlib. SciPy was used to calculate p-values, and a Shapley additive explanations (SHAP) plot was utilized for the model explanation process. The hyperparameters used for machine learning models are discussed in Table 3.

The statistical importance of a feature's association with the target variable is determined in feature selection using a p-value. The clinical features are classified as highly significant, marginally significant, and

non-significant based on the p-values. The following features are highly significant ($p < 0.001$): tumor type ($p = 3.28e-07$), tumor laterality ($p = 6.69e-08$), grade of tumor ($p = 1.05e-05$), FIGOSTAGEGROUP_IIC ($p = 1.21e-26$). The following features are marginally significant ($0.01 < p < 0.05$): size ($p = 0.040$), FIGOSTAGEGROUP_IIB ($p = 0.013$). The following features are least significant ($p \geq 0.05$): age ($p = 0.242$), FIGOSTAGEGROUP_IC3 ($p = 0.083$).

Table 3. Hyperparameters for machine learning model

Model	Hyperparameters
LR	max_iter, penalty, C, solver
SVM	kernel, C, gamma, probability
KNN	n_neighbors, weights, metric
DT	max_depth, min_samples_split, criterion
Voting classifier	estimators, voting
Bagging classifier	estimator, n_estimators, max_samples, bootstrap
Boosting	n_estimators, learning_rate, base_estimator
Stacking classifier	estimators, final_estimator

4.1. Relationship between tumor stage, size, laterality, and CA-125

The most commonly reported tumor marker is CA-125, which has been linked to a number of histological abnormalities that may aid in early detection. Ovarian cancer is typically detected in the later stages and is undiagnosed in its early stages. CA-125 levels greater than 35 U/ml are regarded to be elevated. Figure 3 infers that the patients in the advanced stage have smaller tumors and high levels of CA-125. This fact would be beneficial to the doctors to check the metastasis condition.

Also, the next parameter, bilateral EOC, had a worse prognosis than unilateral EOC. Additionally, as illustrated in Figure 4, we can see that CA-125 is high for the bilateral EOC. Moreover, serous and endometrioid carcinomas comprise the bulk of the population. Compared to endometrioid carcinoma, serous carcinoma typically exhibits greater aggression and a poorer prognosis. Survival rates are greatly impacted by the fact that it is typically identified at stages that are more advanced. Therefore, a greater CA-125 level is seen for serous carcinoma, which is indicated as type 1 and type 2 indicates serous and endometrioid carcinoma as per Figure 5. The essential features are vital for precise classification and prediction. The results of various classifiers, such as LR, DT, SVM, KNN, stacking, bagging, and voting classifiers, were used for the prediction of ovarian cancer. The results of various classifiers are discussed in sections 4.2 and 4.3.

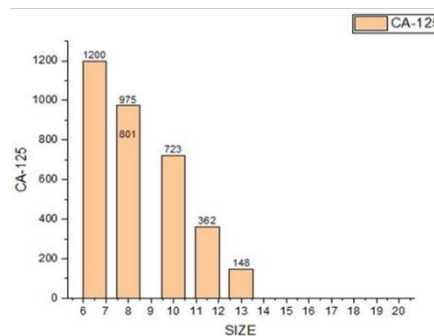


Figure 3. Influence of high level of CA-125 on tumor size

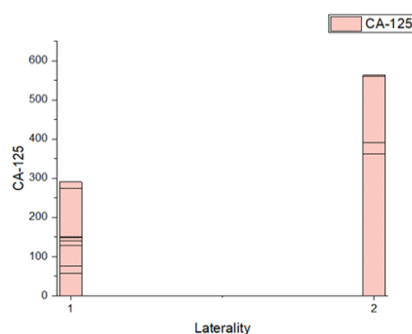


Figure 4. Investigation of laterality with CA-125

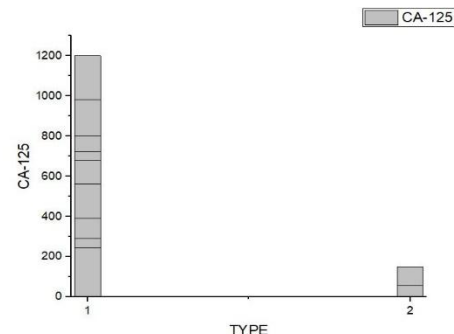


Figure 5. Association of tumor type with CA-125

4.2. Base models

Accuracy, precision, recall, F1-score, sensitivity, and specificity were the primary measures used in the performance evaluation. The following are the performance results that were attained. Table 4 lists all the evaluation metrics, it is observed that the LR from the RFE technique outperforms the other feature selection algorithms in terms of accuracy of 96% and other performance parameters. Accuracy, which is measured as the percentage of correctly categorized cases in a dataset, was the primary evaluation metric employed in this review procedure.

Table 4. Evaluation metrics of base models

Feature selection technique	Base models	Accuracy	Precision	Recall	F1-score	Sensitivity	Specificity
RFE technique	SVM	92	100	85	92.1	85.83	100
	KNN	78.6	94.2	67.5	78.17	67.5	93.81
	DT	94.6	95.28	93.06	95.28	95.28	93.81
	LR	96	100	93.06	96.31	93.06	100
Correlation coefficient	SVM	88	95	83	88.7	83.6	93.81
	KNN	92	93	93	92.9	93.06	90.95
	DT	93.33	95.56	93.06	94.12	93.06	94.29
	LR	88	95.28	83.61	88.74	83.61	93.81

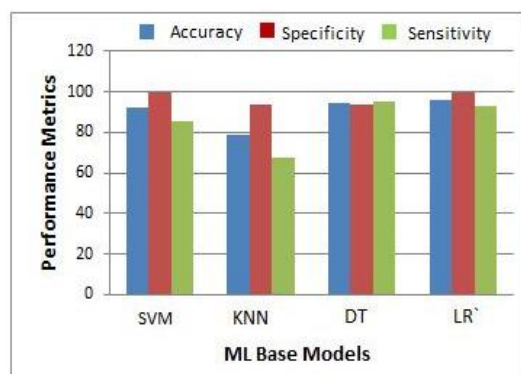
4.3. Ensemble models

There is a noticeable improvement in accuracy, precision, recall, F1-score, sensitivity, and specificity when compared to the base models. Notably, the boosting model shows the highest accuracy of 96% from the RFE technique. Table 5 shows all the evaluation metrics scores of the ensemble model. The aim of this investigation is to determine the optimal and most efficient approach for the detection of ovarian cancer tumors, as well as the elements that contribute to the superior performance of a specific ensemble learning strategy over others.

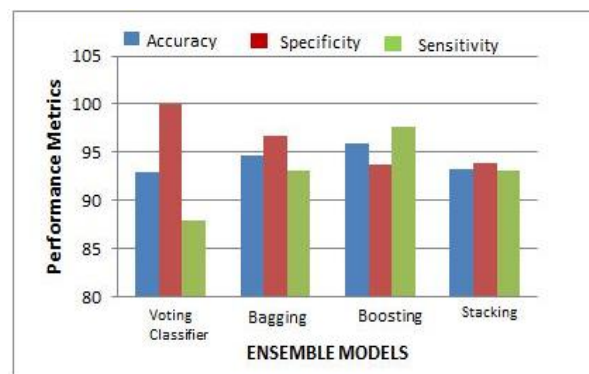
AI-based machine learning algorithms are employed for the prognostic and diagnostic assessment of cancers. Figure 6 presents the comparative performance of different models used in this study. As illustrated in Figure 6(a), base models have a sensitivity of 93.06% and specificity of 100%, while Figure 6(b) demonstrates that ensemble models have a sensitivity of 97.7% and specificity of 93.85% with an accuracy of 96%.

Table 5. Accuracy of ensemble models

Feature selection technique	Base models	Accuracy	Precision	Recall	F1-score	Sensitivity	Specificity
RFE technique	Voting classifier	93	100	88	93.46	88	100
	Bagging	94.6	97.7	93.06	95.27	93.06	96.67
	Boosting	96	95.56	97.7	96.6	98	94
	Stacking	93.3	95.2	93.06	94.1	93.06	93.81
Correlation coefficient	Voting classifier	88	95.2	83.6	88.7	83.6	93.8
	Bagging	89.3	95.2	88.6	91.7	88.6	90.9
	Boosting	93.3	95.5	93.06	94.12	93.06	94.29
	Stacking	92	93.3	93.06	93.07	93.06	90.9



(a)



(b)

Figure 6. Performance evaluation for RFE techniques of (a) base models and (b) ensemble models

4.4. Model interpretations

XAI facilitates the visualization of whether a certain feature is linked to a model's predictions. The model projections are displayed on the y-axis, while the value distribution of the feature is displayed on the x-axis. Based on the SHAP plot, positive values represent higher prediction risk, whereas negative value indicates a decrease in the prediction risk. The SHAP plot, as shown in Figure 7, provides an overview of the features influencing the model's prediction. Figures 7(a) and 7(b) indicate that the tumor size, laterality, and FIGO third stage (IIIC) have the greatest influence on the prediction and are most significant in relation to CA-125. The model's predicted accuracy will be degraded to a greater or lesser degree based on age towards the left.

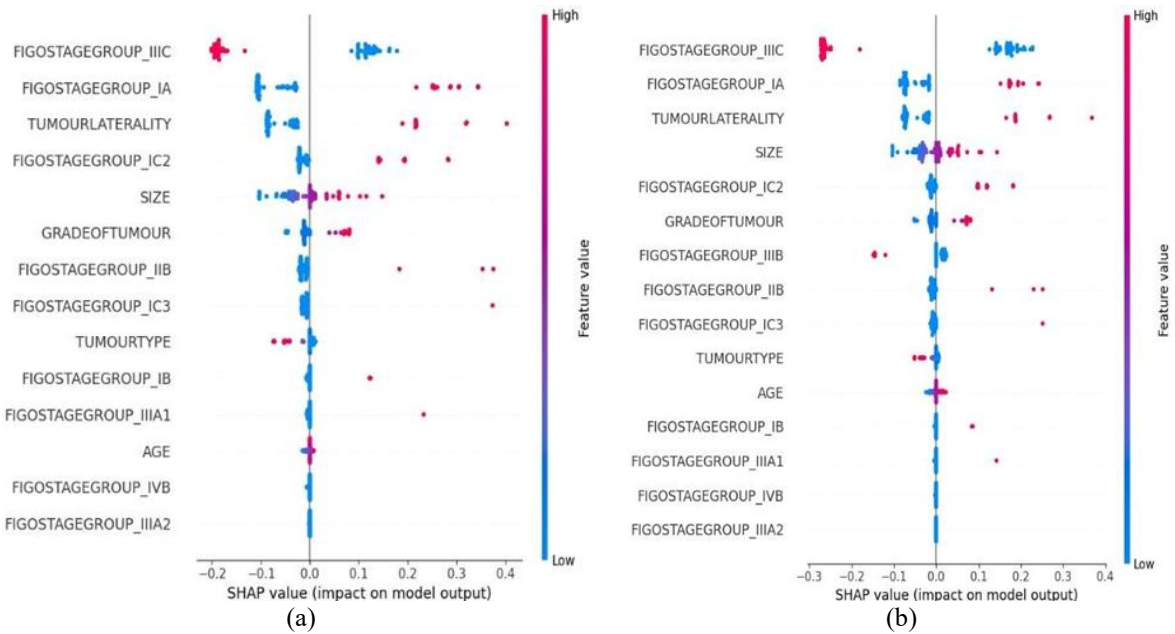


Figure 7. SHAP plot for (a) correlation coefficient and (b) RFE

4.5. Comparative analysis with the existing results

Table 6 presents a comparison between the proposed strategy for medical dataset diagnosis and previous investigations. The results show that the proposed method outperforms earlier studies. The accuracy of prediction is the highest for the proposed work, reaching 96%.

Table 6. Comparison with existing results

Ref	No. of selected features	Methods	Accuracy
[25]	8 features (retrospective study)	GBM, SVM, RF, CRF, NB, neural network, and Elastic Net	Highest accuracy was obtained from RF with 69%
[14]	Top 18 features were selected from Kaggle database	KNN, SVM, DT, and ensemble learning techniques such as max voting, boosting, bagging, and stacking.	Highest accuracy with base model-85% and ensemble model -89%
Proposed Work	Top 7 features were selected (Ramaiah Medical College)	KNN, SVM, and DT along with ensemble models such as max voting, boosting, bagging, and stacking	Highest accuracy with base model-96% and ensemble model-96%

4.6. Discussion

One of the most difficult pathological types is EOC; treatment choices diverge depending on the type of tumor. But in practice, failing to forecast the type of tumor often leads to either a poorly planned operation or a misdiagnosis of the pathological type, both of which can have a substantial impact on the patient. Ovarian malignancies must therefore be diagnosed as soon as possible.

This study intends to determine the best feature selection technique for predicting the preoperative diagnosis of ovarian cancer, by merging it with machine learning algorithms. Among the many facets of ovarian cancer are several histotypes with varying grades and clinical stages. Therefore, a crucial tactic for

delivering individualized, optimal healthcare is predicting the clinical characteristics of ovarian cancer using preoperative data and categorizing patients based on prognosis.

According to Table 6, the majority of research was conducted utilizing machine learning approaches to perform classification on biomarkers alone, such as age, CA-125, menopause, HE-4, and NEU [14], [25]. In addition to classification, determining the preoperative status is the primary goal, which is contingent upon the several forms of EOC, including serous, endometrioid, clear cell, and mucinous carcinoma. These forms also impact in prognosis as well. So, CA-125 is considered the gold tumor marker, which has dependencies on various other clinical factors [22], [23]. Therefore, in this work, CA-125 readings are considered along with several other features, such as tumor size, laterality, FIGO stage, and tumor type.

The dataset used in this study makes it clear that patients in stages of cancer II, III, and IV had smaller tumors and significantly higher CA-125 levels than those in stage I, which is in the metastatic condition. This information on staging assists the doctors in deciding further treatment, such as surgery, chemotherapy, and other treatments, which also helps to analyze the aggressiveness of the cancer. These findings further emphasize the need to consider both the bilateral and unilateral elements, with bilaterality being more commonly seen in those who are at the evolved stage. This data about the origin of the tumor and its potential for spread can be inferred from its size and laterality, or whether it is on either or both of the ovaries. Knowing the type of tumor, aids physicians in determining the best course of therapy and forecasting the patient's prognosis. The dependencies of the above parameters with CA-125 readings are evident from Figures 3 to 5.

Therefore, using effective feature selection procedures, this work takes into account every potential parameter for an effective diagnosis and prognosis of ovarian cancer, including age, CA-125, tumor laterality, size, tumor type, grade of tumor, and FIGO stage. The best accuracy of 96% with sensitivity 93% and specificity 100% for LR and 96% with sensitivity 98% and specificity 94% for boosting classifiers under the RFE technique were projected by ensemble classifiers in conjunction with base machine learning classifiers. As per Table 6, this study achieves the highest accuracy of 96% in predicting the pre-operative analysis of ovarian cancer compared to other previous studies. These results, however, imply that AI could offer useful preoperative biomarker based diagnostic data, enabling a customized treatment plan prior to the main clinical approach in EOC.

5. CONCLUSION AND FUTURE WORK

Early detection can improve the prognosis of ovarian cancer, an aggressive condition. Therefore, in order to deliver personalized, optimal healthcare, it is essential to use preoperative data to anticipate the clinical characteristics of ovarian cancer and to classify patients according to their prognosis. This research uses real-time data from Ramaiah Medical College in Bangalore to investigate the application of pre-operative analysis in the successful detection of ovarian cancer. There are 150 patient records in the databases, and each one has 28 features. The dataset consists of 66 records for non-cancerous patients and 84 records for cancerous patients. Seven most significant characteristics from the dataset age, CA-125, tumor laterality, size, tumor type, grade of tumor, and FIGO stage were chosen using two feature selection methods: RFE and correlation coefficient approach. To get the best result in forecasting the pre-operative state, machine learning algorithms are used in conjunction with ensemble models, including voting classifiers, stacking, bagging, and boosting. Our test findings show that, for LR, we obtained 96% accuracy for the base model and 96% accuracy for the boosting ensemble for the RFE feature selection technique. With the aid of XAI, this analysis offers valuable insights into feature selection, model complexity, and accuracy, offering guidance for enhancing machine learning methods utilized in cancer outcome prediction decision-making. Since histology is regarded as the gold standard in the diagnosis of ovarian cancer, this work can be further enhanced by integrating histopathology images with biomarkers for the same individuals. Additionally, to improve the classification accuracy, a bespoke model can be created using an FPGA and a GPU-based system. As a result, EOC can be detected and classified more quickly, potentially extending the patients' lifetime.

ACKNOWLEDGEMENT

The authors wish to express gratitude to Dr. Mangala Gouri S. R., Prof. and Head, Department of Pathology, Ramaiah Medical College, Bangalore, for providing valuable suggestions to conduct this research work.

FUNDING INFORMATION

This research was not funded by any agency.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Suma Palani Subramanya	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓	
Suma Kuncha	✓			✓						✓	✓	✓	✓	
Venkatapathiah														

C : Conceptualization	I : Investigation	Vi : Visualization
M : Methodology	R : Resources	Su : Supervision
So : Software	D : Data Curation	P : Project administration
Va : Validation	O : Writing - Original Draft	Fu : Funding acquisition
Fo : Formal analysis	E : Writing - Review & Editing	

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no competing interests in the conduct of this study.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [SPS], upon reasonable request.

REFERENCES

[1] K. Sathishkumar, M. Chaturvedi, P. Das, S. Stephen, and P. Mathur, "Cancer incidence estimates for 2022 & projection for 2025: result from National Cancer Registry Programme, India," *Indian Journal of Medical Research*, vol. 156, no. 4, pp. 598–607, 2022, doi: 10.4103/ijmr.ijmr_1821_22.

[2] A. Laios, A. Gryparis, D. Dejong, R. Hutson, G. Theophilou, and C. Leach, "Predicting complete cytoreduction for advanced ovarian cancer patients using nearest-neighbor models," *Journal of Ovarian Research*, vol. 13, no. 1, 2020, doi: 10.1186/s13048-020-00700-0.

[3] K. V. Suma, C. S. Sonali, B. S. Chinmayi, B. J. Kiran, and M. Easa, "CNN models comparison for lung cancer classification using CT and PET scans," *IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, 2022, doi: 10.1109/MysuruCon55714.2022.9972704.

[4] L. Zhang, J. Huang, and L. Liu, "Improved deep learning network based in combination with cost-sensitive learning for early detection of ovarian cancer in color ultrasound detecting system," *Journal of Medical Systems*, vol. 43, no. 8, 2019, doi: 10.1007/s10916-019-1356-8.

[5] S. Hatamikia *et al.*, "Ovarian cancer beyond imaging: integration of AI and multiomics biomarkers," *European Radiology Experimental*, vol. 7, no. 1, 2023, doi: 10.1186/s41747-023-00364-7.

[6] C. Sahli, T. T. Pham, and Kenry, "Machine-learning-assisted analysis of patient clinical biomarkers to improve ovarian cancer diagnosis," *Precision Chemistry*, 2025, doi: 10.1021/prechem.5c00028.

[7] K. R. Kasture, D. Choudhari, and P. N. Matte, "Prediction and classification of ovarian cancer using enhanced deep convolutional neural network," *International Journal of Engineering Trends and Technology*, vol. 70, no. 3, pp. 310–318, 2022, doi: 10.14445/22315381/IJETT-V70I3P235.

[8] J. Lipkova *et al.*, "Artificial intelligence for multimodal data integration in oncology," *Cancer cell*, vol. 40, no. 10, pp. 1095–1110, 2022, doi: 10.1016/j.ccell.2022.09.012.

[9] M. Mathur, V. Jindal, and G. Wadhwa, "Detecting malignancy of ovarian tumour using convolutional neural network: a review," *6th International Conference on Parallel, Distributed and Grid Computing*, pp. 351–356, 2020, doi: 10.1109/PDGC50313.2020.9315791.

[10] Y. Liu, B. C. Lawson, X. Huang, B. M. Broom, and J. N. Weinstein, "Prediction of ovarian cancer response to therapy based on deep learning analysis of histopathology images," *Cancers*, vol. 15, no. 16, 2023, doi: 10.3390/cancers15164044.

[11] K. Chen *et al.*, "Integration and interplay of machine learning and bioinformatics approach to identify genetic interaction related to ovarian cancer chemoresistance," *Briefings in Bioinformatics*, vol. 22, no. 6, 2021, doi: 10.1093/bib/bbab100.

[12] F. Arezzo *et al.*, "A machine learning approach applied to gynecological ultrasound to predict progression-free survival in ovarian cancer patients," *Archives of Gynecology and Obstetrics*, vol. 306, no. 6, pp. 2143–2154, 2022, doi: 10.1007/s00404-022-06578-1.

[13] B. Ziyambe *et al.*, "A deep learning framework for the prediction and diagnosis of ovarian cancer in pre-and post-menopausal women," *Diagnostics*, vol. 13, no. 10, 2023, doi: 10.3390/diagnostics13101703.

[14] J. M. S. Lavanya and S. Pasupathi, "Innovative approach towards early prediction of ovarian cancer: machine learning-enabled XAI techniques," *Heliyon*, vol. 10, no. 9, 2024, doi: 10.1016/j.heliyon.2024.e29197.

[15] V. V. P. Wibowo, Z. Rustam, S. Hartini, F. Maulidina, I. Wirasati, and W. Sadewo, "Ovarian cancer classification using K-nearest neighbor and support vector machine," *Journal of Physics: Conference Series*, vol. 1821, no. 1, 2021, doi: 10.1088/1742-6596/1821/1/012007.

[16] M. M. Ahamad *et al.*, "Early-stage detection of ovarian cancer based on clinical data using machine learning approaches," *Journal of Personalized Medicine*, vol. 12, no. 8, 2022, doi: 10.3390/jpm12081211.




[17] N. Taleb, S. Mehmood, M. Zubair, I. Naseer, B. Mago, and M. U. Nasir, "Ovary cancer diagnosing empowered with machine learning," *International Conference on Business Analytics for Technology and Security*, 2022, doi: 10.1109/ICBATSS4253.2022.9759010.

[18] J. Wang, J. Gao, H. Yao, Z. Wu, M. Wang, and J. Qi, "Diagnostic accuracy of serum HE4, CA125 and ROMA in patients with ovarian cancer: a meta-analysis," *Tumor Biology*, vol. 35, no. 6, pp. 6127–6138, 2014, doi: 10.1007/s13277-014-1811-6.




- [19] R. C. Bast *et al.*, “Biomarkers and strategies for early detection of ovarian cancer,” *Cancer Epidemiology Biomarkers and Prevention*, vol. 29, no. 12, pp. 2504–2512, 2020, doi: 10.1158/1055-9965.EPI-20-1057.
- [20] H. J. Song, E. S. Yang, J. D. Kim, C. Y. Park, Y. S. Kim, and M. S. Kyung, “Improving performance for classifying ovarian cancer with menopause information,” *Proceedings of 4th IEEE International Conference on Applied System Innovation*, pp. 1222–1223, 2018, doi: 10.1109/ICASI.2018.8394509.
- [21] N. Abuzinadah *et al.*, “Improved prediction of ovarian cancer using ensemble classifier and shaply explainable AI,” *Cancers*, vol. 15, no. 24, 2023, doi: 10.3390/cancers15245793.
- [22] E. S. Paik *et al.*, “Prediction of survival outcomes in patients with epithelial ovarian cancer using machine learning methods,” *Journal of Gynecologic Oncology*, vol. 30, no. 4, 2019, doi: 10.3802/jgo.2019.30.e65.
- [23] C. Nayak, A. Tripathy, M. Parhi, and S. K. Barisal, “Early stage ovarian cancer prediction using machine learning,” *International Conference in Advances in Power, Signal, and Information Technology*, pp. 603–608, 2023, doi: 10.1109/APSIT58554.2023.10201764.
- [24] M. Aditya, I. Amrita, A. Kodipalli, and R. J. Martis, “Ovarian cancer detection and classification using machine learning,” *5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques*, pp. 279–282, 2021, doi: 10.1109/ICEECOT52851.2021.9707954.
- [25] E. Kawakami *et al.*, “Application of artificial intelligence for preoperative diagnostic and prognostic prediction in epithelial ovarian cancer based on blood biomarkers,” *Clinical Cancer Research*, vol. 25, no. 10, pp. 3006–3015, 2019, doi: 10.1158/1078-0432.CCR-18-3378.
- [26] K. N. Kang, E. Y. Koh, J. Y. Jang, and C. W. Kim, “Multiple biomarkers are more accurate than a combination of carbohydrate antigen 125 and human epididymis protein 4 for ovarian cancer screening,” *Obstetrics and Gynecology Science*, vol. 65, no. 4, pp. 346–354, 2022, doi: 10.5468/ogs.22017.
- [27] M. J. Alam, J. E. Anawar, K. M. M. Uddin, M. H. Rahman, and M. M. Rahman, “Machine learning techniques for predicting ovarian cancer in its early stages using biomarkers,” *6th International Conference on Electrical Engineering and Information and Communication Technology*, pp. 1257–1262, 2024, doi: 10.1109/ICEEICT62016.2024.10534382.
- [28] V. Saravanan, V. Sankaradass, M. Shanmathi, J. P. Bhimavarapu, M. Deivakani, and S. Ramasamy, “An early detection of ovarian cancer and the accurate spreading range in human body by using deep medical learning model,” *International Conference on Disruptive Technologies*, pp. 68–72, 2023, doi: 10.1109/ICDT57929.2023.10151103.
- [29] F. H. Juwono, W. K. Wong, H. T. Pek, S. Sivakumar, and D. D. Acula, “Ovarian cancer detection using optimized machine learning models with adaptive differential evolution,” *Biomedical Signal Processing and Control*, vol. 77, 2022, doi: 10.1016/j.bspc.2022.103785.
- [30] A. Kumar, R. Sushil, and A. K. Tiwari, “Machine learning based approaches for cancer prediction: a survey,” *SSRN Electronic Journal*, 2019, doi: 10.2139/ssrn.3350294.
- [31] K. V. Suma, S. S. Selvi, P. Nanda, M. Shetty, M. Vikas, and K. Awasthi, “Deep learning approach to nailfold capillaroscopy based diabetes mellitus detection,” *International Journal of Online and Biomedical Engineering*, vol. 18, no. 6, pp. 95–109, 2022, doi: 10.3991/ijoe.v18i06.27385.
- [32] M. Lu *et al.*, “Using machine learning to predict ovarian cancer,” *International Journal of Medical Informatics*, vol. 141, 2020, doi: 10.1016/j.ijmedinf.2020.104195.

BIOGRAPHIES OF AUTHORS



Suma Palani Subramanya    is working as an Assistant Professor in the Department of Electronics and Communication Engineering, East West College of Engineering, Bangalore, India. She acquired an M.Tech. in 2019 from Ramaiah Institute of Technology in VLSI Design and Embedded Systems. She is presently pursuing Ph.D. in Bio-Medical Signal Processing at Ramaiah Institute of Technology, affiliated to Visvesvaraya Technological University. Her areas of interest are artificial intelligence, biomedical signal processing, and VLSI design. She can be contacted at email: sumap1994@gmail.com.



Dr. Suma Kuncha Venkatapathiah    is working as an Associate Professor in the Department of Electronics and Communication Engineering, Ramaiah Institute of Technology, Bangalore, India. She completed her Ph.D. in 2019 from Visvesvaraya Technological University. She is a senior member of IEEE, a fellow of IETE, and a member of IAENG. Her areas of interest are biomedical signal/image processing, embedded system design, and artificial intelligence. She can be contacted at email: sumakv@msrit.edu.