

A Multi-Objective GAPO-Optimized LightGBM Framework for Interpretable Ovarian Cancer Prediction

Chinmayee Nayak

Siksha O Anusandhan University

Alakananda Tripathy

alakanandatripathy@soa.ac.in

Siksha O Anusandhan University

Manoranjan Parhi

Siksha O Anusandhan University

Research Article

Keywords: Ovarian cancer, Feature selection, GAPO, Enhanced LightGBM, Uncertainty analysis, Explainable artificial intelligence, SHAP

Posted Date: October 7th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10776808/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Multi-Objective GAPO-Optimized LightGBM Framework for Interpretable Ovarian Cancer Prediction

Chinmayee Nayak^a, Alakananda Tripathy^a, Manoranjan Parhi^b

^aCentre for Artificial Intelligence and Machine Learning, Siksha 'O' Anushandhan, (Deemed to be University), Bhubaneswar, 751030, Odisha, India

^bCentre for Data Science, Department of Computer Science and Engineering, Siksha 'O' Anushandhan, (Deemed to be University), Bhubaneswar, 751030, Odisha, India

Abstract

Reliable predictive models for ovarian cancer are crucial for early diagnosis and improved clinical decision making. Traditional diagnostic methods frequently lack sensitivity and specificity, requiring advanced machine learning approaches for better results. In this study, a novel model is developed based on the combination of a multi-objective genetic algorithm with pareto optimization and an enhanced light gradient boosting model, referred as GAPO-Enhanced LightGBM, to improve the prediction of ovarian cancer. Human epididymis protein 4, cancer antigen 125, carcinoembryonic antigen, neutrophil count, alphafeto protein, aspartate aminotransferase, lymphocyte count, mean platelet volume, globulin, and menopause were found to emerge through GAPO-driven feature selection. The performance of GAPO-Enhanced LightGBM surpasses traditional models, accomplishing an accuracy of 0.92, precision of 0.90, recall of 0.96, specificity of 0.88 and F1 score 0.93. Uncertainty analysis reveals that feature selection significantly impacts model performance, with AUC decreasing from 0.98 to 0.74 as key biomarkers are excluded. An explainable artificial intelligence technique, shapley additive explanation is used to make feature selection more understandable. The findings establish GAPO-Enhanced LightGBM as a scalable, interpretable and

*Corresponding author: alkanandatripathy@soa.ac.in

Email addresses: chinmayeenayak@soa.ac.in (Chinmayee Nayak), alakanandatripathy@soa.ac.in (Alakananda Tripathy), manoranjanparhi@soa.ac.in (Manoranjan Parhi)

clinically viable artificial intelligence-driven framework to assess the risk of ovarian cancer, supporting early intervention and personalized treatment strategies.

Keywords: Ovarian cancer, Feature selection, GAPO, Enhanced LightGBM, Uncertainty analysis, Explainable artificial intelligence, SHAP

1. Introduction

Ovarian Cancer (OC) remains one of the most antagonistic and fast-spreading gynecological cancer, characterized by its asymptomatic advancement and lack of feasible early screening [20]. The implementation of traditional diagnostic methods means that a lot of cases can be diagnosed only at advanced stages. This implies that if the disease is diagnosed late, the mortality rates are much higher and the prognosis usually worse. Traditional diagnostic methods have proved to be useful when it comes to data gathering, though they are usually not effective enough for diagnosing the disease and reacting to it in due time. Due to the nature of ovarian cancer, which includes many genetic, biochemical and medical factors, there is a need for better models that could work on prediction of the disease [32, 44].

Machine Learning (ML) has become an essential tool in predicting cancer treatment by interpreting complicated data sets and recognizing patterns normally overlooked by traditional statistics [11, 1]. Therefore, a model based on ML can serve to diagnose tumors more accurately, assist in early detection, and apply personalized treatment methods, even though current algorithms are still subject to serious problems concerning feature selection, costs of computations [6], and fitting the model too closely to the specific data. Thus, it is necessary to use good optimization methods for feature selection and improving the predictive efficiency.

This study represents a multi-objective evolutionary process for OC prediction by integrating a Multi-Objective Genetic Algorithm with Pareto Optimization (GAPO) and Enhanced Light Gradient Boosting Machine (Enhanced LightGBM). The proposed methodology is presented to improve predictive performance by systematically

providing the most important clinical and biochemical features improving model comprehensibility. In contrast to conventional ML models that require lengthy feature selection methods [25], GAPO uses evolutionary optimization approaches to balance feature reduction with diagnostic consistency, providing precise prediction process.

The Enhanced LightGBM model was selected due to its remarkable efficiency and scalability in managing data having larger number of features. Compared to conventional gradient boosting techniques, Enhanced LightGBM adopts a leaf-wise growth strategy, resulting in faster training times and better performance without reducing accuracy [27]. Furthermore, it basically manages unbalanced datasets, which is especially crucial for clinical applications since the distribution of positive instances is often less prevalent than negative cases [46]. Combining GAPO with Enhanced LightGBM, the proposed model (GAPO-Enhanced LightGBM) provides an optimal balance between computational cost, predictive accuracy, and model comprehensibility, making it well-suited for OC prediction. The data set was endured to extensive preprocessing steps, including missing value imputation, normalization, and outlier detection, to ensure data integrity. Feature selection was implemented using GAPO, which determined the most important biomarkers that influence OC prediction, lowering dimensionality while maintaining diagnostic precision.

The efficacy of the GAPO-Enhanced LightGBM model was evaluated against the existing machine learning models such as Decision Tree (DT) [17], Extreme Gradient Boost (XGBoost) [13], Adaptive Boosting (AdaBoost) [53], Extra Trees Random Forest (Extra Trees RF) [47], Linear Support Vector Machine (Linear SVM) [7], Quadratic SVM [18], and Cubic SVM [18] with the usage of GAPO-based feature selection. The comparative analysis was conducted based on the performance metrics like accuracy, recall, precision, specificity, and F1 score. The comparative analysis indicated that the GAPO-Enhanced LightGBM model has been capable of demonstrating high predictive accuracy. It confirms that the proposed technique of evolutionary optimization-based

selection is an effective tool in clinical decision-making. Apart from the effectiveness of the model in prediction, it has significant prospects in early detection and personalized treatment of OC. By applying appropriate selection of biomarkers and demographic indicators, the model can provide assistance to doctors in assessing risks, planning treatment, and conducting monitoring of patients. Furthermore, the combination of multi-objective optimization approaches assures that the model retains flexibility and generalizable across various clinical datasets, resolving significant challenges in model interpretability and stability in medical Artificial Intelligence (AI) research.

In comparison with other ML models, the efficacy of the GAPO-Enhanced LightGBM model is examined such as Decision Tree (DT), Extreme Gradient Boost (XGBoost), Adaptive Boosting (AdaBoost), Extra Tree Random Forest (Extra Tree RF). The use of both evolutionary optimization and advanced ML methods proves to be useful in OCA prediction making it adaptable and clinically beneficial. The research proves the practice of combination of multi-objectives evolutionary mechanisms for quick diagnostics and elimination of misclassifications. As the field of AI in healthcare continues to progress, such framework can facilitate to refine more accurate, precise, and patient-centric cancer diagnosis systems, ultimately improving better clinical outcomes. This study not only highlights the model reliability it also aims to enhance the transparency in decision-making process using an Explainable Artificial Intelligence (XAI) technique i.e., Shapley Additive Explanation (SHAP).

The primary goal of this article is to predict OC possibility using GAPO-Enhanced LightGBM. The study focuses on the following:

- Implementation of GAPO to select clinically significant cost-effective features that identify the occurrence of OC.
- The Enhanced LightGBM model is used to improve the exactness and accessibility of OC prediction.
- The XAI technique, i.e. SHAP is utilized for enhancing the interpretability of

the feature importance.

The following sections of this article are organized as follows: Section 2 presents an extensive analysis of recent research studies exploring various ML models for OC detection. Section 3 delves into the specific design of the proposed model with a concise overview of the methodologies. Section 4 presents the obtained outcomes from the proposed model, along with a thorough comparative analysis with existing models. Additionally, it discusses uncertainty analysis of the features and utilizes XAI techniques such as SHAP, to provide a more interpretable understanding of feature importance. Finally, Section 5 concludes the important insights of the article with future work.

2. Related Work

OC is a highly invasive cancer with major public health consequences, exacerbated by the lack of viable surgical options [19]. Nopour [38] investigated the application of ML to predict the likelihood of OC occurrence and improve preventive efforts. Six ML models, XGBoost, Random Forest (RF), J-48 DT, SVM, K-Nearest Neighbor (KNN), and Artificial Neural Network (ANN) were deployed. The Area Under the Receiver Operating Characteristic Curve (AUROC) was leveraged to assess the result obtained from employed models performed better than other models with an AUROC of 0.93 (95% Confidence Interval (CI): [0.91–0.95]). However, limitations include the use of data from a particular region, insufficient information on feature selection, and the possible influence of biases or missing data, highlight the necessity for additional validation with diverse datasets and transparent approaches.

According to Sun *et al.* [48], the study employed data from 713 patients from Sun Yat Sen Memorial Hospital, taking into consideration 10 preoperative peripheral blood markers together with their respective ages. Four different machine learning models that

were used include Random Forest (RF), XGBoost, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Logistic Regression (LR). The research revealed that Random Forest was able to achieve the best AUROC value of 0.86 in distinguishing malignant tumors, while the logistic regression model was the best model in terms of the AUROC value of 0.95 in predicting prognosis of the pathological tissue. Further validation was carried out with data from the Guangdong Maternal and Child Health Center that revealed 78.9% and 71.9% accuracy for Random Forest and Logistic regression respectively. However, the study had some limitations because of using only one institution for data collection, limited explanation of feature selection robustness and limited generality of this model and thus requiring further validation.

The precise identification of Benign Ovarian Tumors (BOTs) in comparison to Ovarian Cancers (OCs) is vital for appropriate clinical management of patients. Lu *et al.* [30] used data from 349 Chinese patients with 49 clinical markers to apply the mutually information Maximum Relevance Minimum Redundancy (MRMR) feature selection method with which eight key features were selected among 235 patients (89 BOTs, 146 OCs). A DT model based on two biomarkers human epididymis protein 4 (HE4) and carcinoembryonic antigen (CEA) was developed and tested on 114 patients (89 BOT, 25 OC). The model of the risk of ovarian malignancy algorithm performed better and while LR offered an easy way to perform. However, the dependency on a specific data set limits its generalizability, requiring additional validation.

The effectiveness of adaptive differential evolution was compared by Juwono *et al.* [22]. The weighting characteristics and parameter optimization was examined for the identification of OC. The Least Absolute Shrinkage and Selection Operator (LASSO) was used to reduce the feature subset by reducing irrelevant features while keeping weakly discriminative features. KNN and SVM were used with particular regularization parameter setting. However, reliance on particular regularization techniques and restricted by a single dataset validation restricts the generalizability of the approach. An AI-based

modeling of ovarian tumor biology using data from 202 patients (53 OC, 23 borderline, 126 benign) and 16 preoperative factors from blood tests, demographics and imaging was evaluated by Akazawa *et al.* [3]. Five ML estimators XGBoost, SVM, naive bayes, RF, and LR were compared for assessing the highest precision of 80%. The relevance analysis of characteristics differed between the correlation coefficients, the regression coefficients, and the RF rankings. While AI exhibits promising results, the study moderates accuracy and the lack of external validation emphasizes the need for additional optimization and wider validation.

An ML system to predict the key clinical factors including stage, histotype, and prognosis in Epithelial OC (EOC) was proposed by Kawakami *et al.* [24]. Evaluating 334 patients with EOC and 101 patients with benign OCs using seven ML classifiers, including RF, Conditional RF and Gradient Boosting Machine (GBM). The study also used 32 previous biomarkers and age for the predictions. RF demonstrated better performance, with 92.4% accuracy and an AUC of 0.968 for segmenting EOC from benign cancers, and 69.0% accuracy and an AUC of 0.760 for clinical stage estimation. Moreover, RF also estimated high-grade serous and mucinous histotypes and surgical results. However, Whitwell *et al.* [52] aims to improve the detection of OC by adding additional biomarkers to longitudinal models in addition to CA125. Using 490 successive serum samples from 49 women with OC and 31 with cancer-free status, proteomic-based marker identified candidates including HE4, CHI3L1, PEBP4 and AGR2. The models displayed lower predictive values compared to CA125, suggesting the potential to improve early identification and independent validation.

Zhang *et al.* [54] emphasizes various feature selection techniques, including Principal Component Analysis (PCA), Locally Linear Embedding (LLE), MRMR, LASSO regularization, and biomarker-focused approaches, which maintain dimensionality reduction, interpretability, and model efficiency in the prediction of OC. Although machine learning techniques have been successfully used with multiple applications, certain ob-

stacles such as the lack of transparency in feature selection processes and insufficient validation through diverse populations hinder their generalizability and reproducibility. Legitimately, there are many studies that favor statistics over biological interpretability. For instance, classical methods including PCA and LLE are mainly applied instead of biomarker-based approaches. Concerning challenges, eliminating biases, bridging missing data gaps, and managing variability in collected data, including multimodal and longitudinal data, are essential for obtaining more robust results.

In this recent study [41], three machine learning algorithms (RF, CatBoost, and SVM) along with four feature selection methods (Boruta, PCA, RFE, and Mutual Information) have been assessed recently. The research demonstrated that Boruta and RFE show superior performance in terms of diagnostic accuracy, while RF and Catboost show the highest AUC values according to diagnostic accuracy, these methods show the same performance through different techniques. However a systematic review was conducted by Noei *et al.* [37] on ten trials (n=3,460) evaluated the performance of AI in predicting outcomes following Complete Cytoreduction in OC. AI models outperformed better performance compared to conventional statistics, with aggregated accuracy for overall survival at 68.64% and for R0 reconstruction at 82.5%. The majority of studies concluded that ML and ANNs imposed more accurate predictions, underscoring AI's ability to improve outcome prognosis in postoperative OC therapy.

A systematic review was conducted by Piedimonte *et al.* [42] to explore the application of neural network emerging ML in OC prediction. This review mainly focused on the use of radioactive substances in predicting therapy response, surgical results, and survivability. It mainly focuses on the use of ML and radioactive substances in predicting therapy response, surgical results, and survivability. For anticipating response to neoadjuvant cancer treatment, models accrued performance metrics approximately 80%, leveraging clinical factors and imaging data an enhancement over conventional methods that may aid in surgical candidacy decisions. In surgical procedures, similar

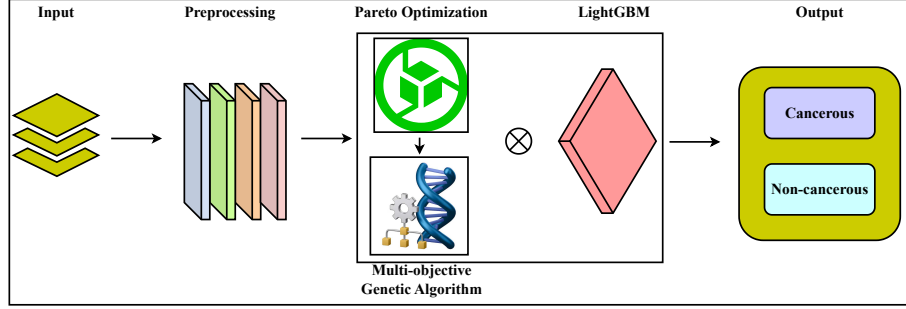


Figure 1: Framework of the proposed GAPO-Enhanced LightGBM model

accuracy rates were achieved, with one model in a platinum-resistant that achieving an AUC more than 95%, suggesting the potential of clinical utility in directing specialized therapies such as mirvetuxumab.

Despite emerging focus in ML systems for the OC detection, there is a lack of a clinically reliable and understandable model. The majority of recent approaches disregarding the uncertainty inherent in clinical features such as overlapping biomarkers hindering to diagnostic reliability. Additionally, comprehensibility and model transparency remain unexplored, restricting clinician trust and implementation. The presented article seeks to overcome the critical gap in existing studies concentrating on the development of clinically reliable and explicable ML models for the OC detection.

3. Methodology

This section provides the study methodology, which includes a detailed description of the data collection, preprocessing steps, and the proposed methodology. Figure 1 illustrates the proposed framework of the study.

3.1. Data sourcing and pre-processing

The ovarian cancer dataset utilized in this study was sourced from Mendeley [30]. The data set comprises a variety of functional biomarkers, physical parameters, and blood chemistry reports from female patients. All histological information of the patients was analyzed according to the standards of World Health Organization (WHO).

The data set includes 349 records and each containing 49 features listed in Table 1. The target variable, labeled 'TYPE', is binary, with '1' indicating a cancer diagnosis and '0' indicating a non-cancerous diagnosis. The data collection process involves aggregating clinical test results that are frequently utilized in the diagnosis of OC. These tests included tests for kidney function, blood tests, cancer biomarker evaluations (CA125, CA74-4, CA19-9), and other physical health indicators (Age, Menopause). The data set was selected for its extensive coverage of features relevant to OC detection. Key features include enzyme levels, proteins, and blood components associated with cancer risk. The selection of the data set is rely on the diversity of features and the presence of clinically meaningful biomarkers.

To ensure reproducibility, detailed attention was given to preprocessing steps, including the handling of missing data, normalization of continuous variables, and encoding of categorical features. Columns with missing values accounting for more than 20% of total entries are removed. Hence, columns 'Mg' have been dropped from the data frame by imputing the null values with the mean of the corresponding features. In the 'MONO', 'HGB' and 'GLU' columns, the mean of the centrality measures was used to impute for missing values. After all the missing value imputation, a completed data set with no missing value resulted with 47 features. After that, standard scalar normalization is used to convert the entire feature into a single or similar scale.

3.2. Feature Selection

Feature selection is a critical process that involves selecting the features of the original set and ensuring that they retain the majority of the statistical information represented by the entire data set [12]. Clinical data sets often include a high number of medical features, which can obscure a meaningful relationship with the target variable and make the ML model computationally intensive. Furthermore, the presence of numerous features can lead to complexity and less interpretable algorithms, making it challenging to extract actionable insights for medical applications [33, 16]. This

Table 1: Statistical Summary of the Dataset

Name of the feature	Feature	Standard Deviation	Standard Error	IQR	p-value
Alphafeto Protein	AFP	355.78	14.84	182.95	0.416
Albumin/Globulin Ratio	AG	6.38	0.27	9.12	0.356
Albumin	ALB	7	0.29	10.66	0.621
Alkaline Phosphate	ALP	186.34	7.77	88.75	0.689
Alanine Transiminase	ALT	21.23	0.89	35.38	0.701
Aspartate Aminotransferase	AST	17.38	0.72	28	0.126
Age	Age	16.82	0.7	26.21	0.953
Count of Basophil	BASO#	0.03	0	0.05	0.338
Percentage of Basophil	BASO%	0.49	0.02	0.77	0.361
Blood Urea Nitrogen	BUN	2.12	0.09	3	0.346
Cancer Antigen 125	CA125	1323.64	55.2	998.88	0.162
Cancer Antigen19-9	CA19-9	259.42	10.82	229.38	0.355
Cancer Antigen72-4	CA72-4	38.56	1.61	18.56	0.199
Carcinoembryonic Antigen	CEA	37.32	1.56	29.16	0.122
Chloride	CL	5.43	0.23	7.41	0.981
Carbon Dioxide and Combining Power	CO2CP	4.11	0.17	6.12	<0.001
Creatinine	CREA	16.72	0.7	24.34	<0.001
Calcium	Ca	0.54	0.02	0.89	<0.001
Bilirubin	DBIL	2.97	0.12	4.92	0.482
Eosinophil Count	EO#	0.11	0	0.19	0.334
Eosinophil percentage	EO%	2.04	0.09	3.07	0.816
Gamma-glutamyl transferase	GGT	47.89	2	75.08	0.214
Globulin	GLO	6.61	0.28	8.1	<0.001
Glucose	GLU.	2.17	0.09	3.38	0.524
Hematocrit	HCT	0.07	0	0.08	<0.001
Human Epididymis Protein	HE4	974.1	40.62	619.22	0.191
Hemoglobin	HGB	25.92	1.08	35.38	<0.001
Indirect Bilirubin	IBIL	6.83	0.28	9.34	0.671
Potassium	K	0.55	0.02	0.79	<0.001
Number of lymphocytes	LYM#	0.8	0.03	1.24	0.778
Percentage of lymphocytes	LYM%	11.78	0.49	18.32	0.128
Mean Corpuscular Hemoglobin	MCH	4.41	0.18	6.57	<0.001
Mean Corpuscular Volume	MCV	10.98	0.46	14.94	<0.001
Mononucleosis Count	MONO#	0.25	0.01	0.39	0.108
Mononucleosis percentage	MONO%	4.43	0.18	5.43	0.514
Mean Platelet Volume	MPV	2.24	0.09	3.36	<0.001
Menopause	-	-	-	-	-
Magnesium	Mg	0.18	0.01	0.26	<0.001
Neutrophil Count	NEU	12.58	0.52	18.4	<0.001
Sodium	Na	5.69	0.24	6.8	<0.001
Procalcitonin	PCT	0.14	0.01	0.17	0.353
Platelet Distribution Width	PDW	3.45	0.14	5.64	<0.001
Phosphorus	PHOS	0.28	0.01	0.41	<0.001
Platelet Count	PLT	183.14	7.64	188.62	0.171
Red Blood Cell	RBC	0.87	0.04	1.03	<0.001
Red Blood Cell Distribution Width	RDW	2.87	0.12	4.61	<0.001
Total Bilirubin	TBIL	8.65	0.36	11.91	0.145
Total Protein	TP	12.93	0.54	16.6	<0.001
Urinalysis	UA	120.35	5.02	167.27	0.307

intricacy not only hampers understanding of key features, but also limits the recognition of the practical usability of predictive models in clinical decision-making [8, 9]. Therefore, effective feature selection is essential to streamline the feature set, improving interpretability and increasing the efficiency of the computational process.

Generally, the filter type feature selection method (correlation, mutual information, chi-squared test and variance threshold method) evaluates the features independently and ignores the interaction (co-linearity) between the features, leading to redundancy among the selected features [10]. The selection of the threshold value in the filter method is inherently arbitrary, which can lead to inconsistency in the resulting subset of features. However, wrapper feature selection methods involve an iterative process of exploring the feature space by training and validating a model on various subsets of feature, which is exhaustive and depends upon the choice of model. Genetic Algorithm (GA) [34] is an evolutionary type of feature selection method based on genetics and natural selection. The GA seeks to select the informative subset of features by iteratively refining and evaluating successive feature subset. GA uses biological operators including selection, crossover and mutation to obtain the refined set of features. The method starts by producing a starting population of potential solutions, where each potential refers to a subset of the features [49, 5]. To create new solutions, known as offspring, these candidates undergo evolutionary changes like recombination and mutation. Each possible solution is represented by a structure called a chromosome, which blends feature subsets.

The algorithm begins with an arbitrarily initialized population of chromosomes [15]. During the evolution process, the new populations are created over successive series. The quality of each candidate in the population is calculated using a fitness function and a fitness score is given to each based on its performance [39]. The selection process identifies the fittest individuals using a stochastic technique such as event selection. Individuals with higher fitness scores are more inclined to be selected as parents for

recombination. Then the crossover operator is used to confirm the variations and continue the exploration of the feature space. It performs a recombination of the parent chromosomes, often using operations like XOR, to create offspring that differ slightly from their parents [50].

To further strengthen the algorithm's inquisitive capabilities, the mutation function is utilized. This introduces unanticipated modifications to the genes encoded by chromosomes, increasing more diversity and reducing the likelihood of early convergence [21]. In addition, a hierarchical selection strategy is implemented to ensure that the top performing individuals from the present generation are retained and propagated to the subsequent generation without alteration. The newly formed population, improved with diverse and potentially superior solutions, is re-evaluated using the objective function. This continuous process continues for several iterations, progressively advancing the population. Continuing this process, GA identifies subsets of features that are more suitable for prediction, ultimately identifying the feature space with a balance of exploration and deployment [36]. GA is useful for feature selection tasks that confront notable setbacks that impair their performance in sensitive scenarios like feature selection in clinical datasets. These include difficulties in managing data with a large number of features, slow convergence, premature stagnation in local optima and excessive computational costs. Furthermore, GAs frequently lack mechanisms for adaptive parameter modification, multi-objective optimization, and fast search space, which are crucial for clinical data with limited samples and a large number of features [45]. To resolve these limitations, GAPO introduce enhancement such as hybridization with localization techniques and ML models to improve solutions along the pareto front, dynamic parameter adjustment to adapt mutation and crossover rates according to the optimization process, and extensive initialization process by domain knowledge to ensure diversity and optimal convergence toward pareto-optimal solutions.

3.2.1. Multi-objective Genetic Algorithm with Pareto Optimization

GAPO is a comprehensive multi-objective evolutionary approach designed to confront the challenges of solving extensive feature optimization problems. The conventional GA focuses on a single objective where GAPO leveraged to maximize several conflicting objectives simultaneously by adopting the pareto front computation [23]. Key advancements in GAPO include integration with the local search technique to improve solutions within the pareto front, dynamic parameter adjustment to alter mutation and crossover rates during the optimization process, and refined population initialization procedures to ensure variation and convergence toward pareto-optimal responses [26].

The GAPO feature selection was performed on the OC dataset represented as $\mathcal{D} = \{(X_{ij}, y)\}$, where $X_{ij} \in \mathbb{R}^{n \times m}$ signifies the feature matrix and y the class variable. Initially, the feature matrix X is standardized, and a population $\mathcal{P}_0$ of p individuals is generated, where every individual $V_k = [c_1, \dots, c_m]$ holds feature relevance levels or weight of the features. Features with relevance levels exceeding a threshold $\beta = 0.5$ are selected per Equation 1. Individuals with no specific selected features are initialized as zero fitness and penalized along with maximum feature count. For qualified individuals, the chosen subset is utilized to train a RF classifier using stratified k -fold cross-validation. Fitness objectives are then utilized to perform non-dominated sorting to produce pareto fronts with respect to dominance relations $F_k(V_i) \leq F_k(V_j)$, subsequently diversity is preserved using crowding distance Equation 2. Selection for recombination is performed on the basis of pareto front rank and higher diversity. New individuals are formed using blend crossover expressed in Equation 3, and mutation by including Gaussian noise in Equation 4. The population is maintained by preserving the top individuals based on pareto dominance and higher crowding distance as the Equation 5. The best individual is chosen from each subsequent generation based on maximal fitness (refer Equation 6) and the associated feature indices are recorded in Equation 7. The algorithm outputs balance a diverse array of solutions to preserve

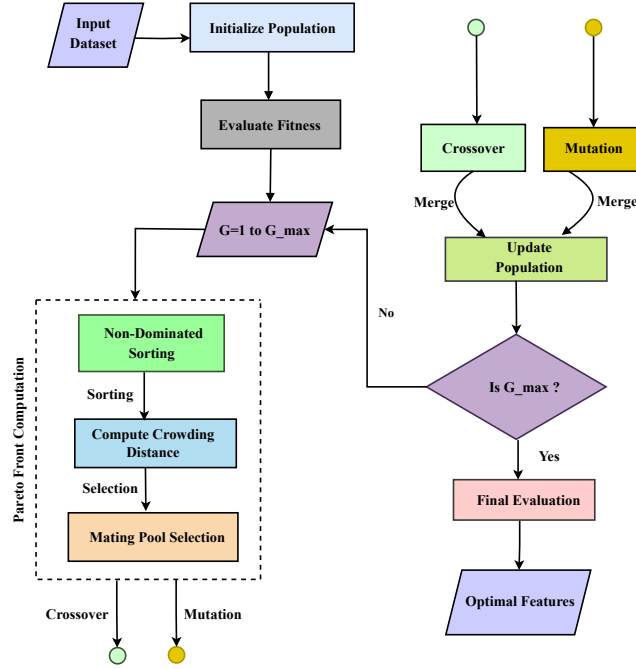


Figure 2: Working Principle of Multi-Objective GAPO

premature convergence and balance conflicting objectives, such as accuracy and computational efficiency, through the significance of a pareto front. Algorithm 1 and Figure 2 presents detailed procedures for feature selection using GAPO.

3.3. Enhanced Light Gradient Boosting Model

Enhanced LightGBM is the gradient boosted DT model basically used for the binary class classification task. It is built to optimize prediction that balances the rigor of the evaluation and computational efficiency. In binary class classification, Enhanced LightGBM predicts probabilities using tree-based learning approaches and sigmoid function [14, 40]. Minimizing log-loss and implementing a suitable regularization term, Enhanced LightGBM shows an accurate and scalable solution for binary classification task [35]. Algorithm 2 and Figure 3 present the overall process of the Enhanced LightGBM model. Equation 8 used an estimated function in Enhanced LightGBM to optimize the aggregate loss across the entire data, the loss function expressed in

Algorithm 1 Multi-Objective Genetic Algorithm using Pareto Optimization

Require: • Original dataset $\mathcal{D} = \{(X_{ij}, y_i) \mid i \in [1, n], j \in [1, m]\}$, where $X_{ij} \in \mathbb{R}^{n \times m}$ is the feature matrix with n samples and m features, and y_i is the target variable.

• Parameters for Genetic Algorithm:

– Population size N , Generations number G , Selection Threshold θ .

- 1: **Initialize** population $|\mathcal{P}_0|=p$ i.e. $\mathcal{P}_0=\{V_1, V_2, \dots, V_p\}$ where $V_k = [c_1, \dots, c_m]$
- 2: **Evaluate** the fitness of each individual $V_i \in \mathcal{P}_0$ for the objectives $F_k(V_i)$, $k = 1, 2, \dots, m$.
- 3: **for** each individual $V_k \in \mathcal{P}_0$ **do**

$$F_k = \{i \mid c_i > 0.5, c_i \in V_k\}. \quad (1)$$

- 4: **if** $|F_k| = 0$ **then**
- 5: $F(V_k) = 0$
- 6: **end if**
- 7: **for** $G = 1$ to $G_{\max}$ **do**
- 8: **Perform** non-dominated sorting on $\mathcal{P}_G$ to generate pareto fronts $\{F_1, F_2, \dots\}$.
- 9: **Sort** individuals as:
- 10: **if** $V_i \prec V_j \iff \forall k, F_k(V_i) \leq F_k(V_j)$ **then**
- 11:

$$d(V_i) = \sum_{k=1}^p \frac{F_k(V_{i+1}) - F_k(V_{i-1})}{\max(F_k) - \min(F_k)} \quad (2)$$

- 12: **end if**
- 13: Pair of parents (V_p, V_q) in the mating pool
- 14: Generate offspring using the blend crossover operator:

$$V_{\text{child}} = \alpha V_p + (1 - \alpha) V_q, \quad \alpha \sim \mathcal{U}(0, 1) \quad (3)$$

- 15: Mutate the offspring weights with Gaussian noise:

$$w'_j = w_j + \delta, \quad \delta \sim \mathcal{N}(0, \sigma), \quad \text{with probability } \mu \quad (4)$$

- 16: **end for**
- 17: $\mathcal{P}_{G+1} = \cup_{i=1}^{k'} F_i \cup \{x \in F_{k'+1} \mid x \text{ has the highest } d(x) \text{ values}\}$ (5)

- 18: **end for**
- 19: Identify the best individual $V_{\text{best}} \in \mathcal{P}_G$ with the maximum fitness score:

$$F_{\text{best}} = \max_{V_k \in \mathcal{P}_G} F(V_k) \quad (6)$$

- 20: Collect the selected feature indices:

$$F_{\text{optimal}} = \{i \mid c_i > 0.5, i \in V_{\text{best}}\} \quad (7)$$

Ensure: Subset of pareto-based optimal features, F_{optimal} .

Equation 9 and to prevent overfitting used regularization term as shown in Equation 10.

$$\begin{aligned}\mathcal{L} &= \sum_{i=1}^n \leq(y_i, y_i^*) + \Omega(\mathcal{U}) \\ &\approx \sum_{i=1}^n \leq(y_i, y_i^{*(t-1)}) + g_i \mathcal{U}(x_i) + \frac{1}{2} h_i \mathcal{U}(x_i)^2 + \Omega(\mathcal{U})\end{aligned}\quad (8)$$

where, $\leq(y_i, y_i^*)$ is the binary cross entropy-loss function used to calculate the deviation between true class level y_i to estimated class level y_i^* which is expressed in Equation 9:

$$\leq(y_i, y_i^*) = -[y_i \log(y_i^*) + (1 - y_i) \log(1 - y_i^*)] \quad (9)$$

and the regularization parameter $\Omega(\mathcal{U})$ is characterized by the following Equation 10

$$\Omega(\mathcal{U}) = \gamma n_{leaves} + \frac{1}{2} \lambda \sum_{j=1}^n \omega_j^2 \quad (10)$$

where,

γ = Regularization parameter, n_{leaves} = number of leaves in the tree, λ = Regularization parameter to penalize the effect of the L2 term, $\frac{1}{2} \sum_{j=1}^n \omega_j^2$ = L2 regularization term.

To minimize the loss function Enhanced LightGBM used second order Tayler series expansion for each data point in the dataset around previous approximation presented by Equation 11

$$\leq(y_i, y_i^*) \approx \sum_{i=1}^n \leq(y_i, y_i^{*(t-1)}) + g_i \mathcal{U}(x_i) + \frac{1}{2} h_i \mathcal{U}(x_i)^2 \quad (11)$$

where,

$\leq(y_i, y_i^{*(t-1)})$ is loss function measured from the previous iteration (t-1), $\mathcal{U}(x_i) = \frac{1}{1+e^{-x_i}}$ is the predicted value of the current tree, g_i is the error gradient function, h_i is the second-order derivative of the loss function.

3.4. Explainable AI

XAI is a set of methods that makes convoluted ML systems more comprehensible to humans [31]. As AI systems decisions become increasingly incorporated into the healthcare field, the ability to understand and recognize their decision is crucial.

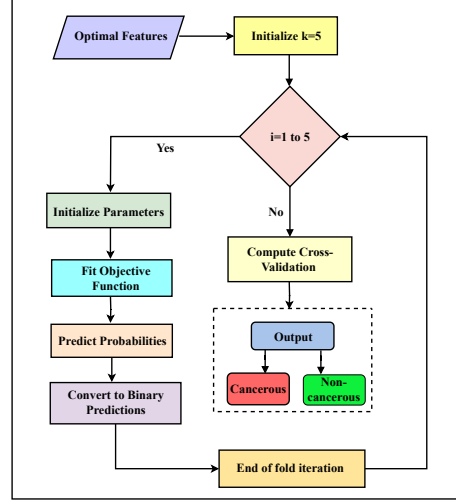


Figure 3: Working Principle of Enhanced LightGBM model

XAI seeks to plug the disparity between model complexity and human understanding, enabling users to trust, comprehend, and successfully control AI systems in crucial decision-making situations. In the context of OC prediction, clinicians and researchers need to understand which features most significantly influence a models output, enabling more informed and sure decision-making. To assure interpretability in the OC prediction, SHAP was used. It provides a mathematically valid and systematic approach for determining each feature contribution to the models prediction. It relies on the concept of Shapley standard from multilayer game theory, which ensures equitably distribution of value among features. The Shapley standard for a feature can be calculated by Equation 14.

$$\Phi_i(\mathcal{U}) = \sum_{\mathbb{S} \subseteq \mathcal{N} \setminus \{i\}} \frac{|\mathbb{S}|!(|\mathcal{N}| - |\mathbb{S}| - 1)!}{|\mathcal{N}|!} [\mathcal{U}(\mathbb{S} \cup \{i\}) - \mathcal{U}(\mathbb{S})] \quad (14)$$

The Shapley standard $\Phi_i(\mathcal{U})$ quantifies the proportion of contribution of particular feature to the prediction, where i represents a particular feature in the model $\mathcal{U}$. $\mathcal{N}$ presents the full set of features and $\mathbb{S}$ presents the subset of features. $\mathcal{U}(\mathbb{S})$ is the model prediction on the selected feature subset $\mathbb{S}$. The expression $\mathcal{U}(\mathbb{S} \cup \{i\}) - \mathcal{U}(\mathbb{S})$

Algorithm 2 Enhanced Light Gradient Boosting Model

Require: Dataset $\mathcal{D} = \{(X = X_{ij}), y = y_{ij}\}$, where $X_{ij} \in \mathbb{R}^{n \times m}$ is the feature matrix and $y_{ij} \in \{0, 1\}^n$ is the binary target variable, Parameters for Enhanced LightGBM: objective=Binary, boosting_type=gbdt, convergence rate parameter= η , total estimators=n_estimators, number of leaves=num_leaves, random_state

- 1: **Initialize** $k = 5$
- 2: **for** $i = 1 : k$ **do**
- 3: Split data:
 $X_{\text{train}}, X_{\text{test}}, y_{\text{train}}, y_{\text{test}} = \text{split}(X, y, p)$
- 4: Initialize Enhanced LightGBM classifier with parameters.
- 5: Train the model using the Enhanced LightGBM objective function expressed in Equation 8
- 6: Predict probabilities as Equation 12:

$$y_i^* = \sigma(\mathbf{U}(X_{\text{test}})) \quad (12)$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function.

- 7: Convert probabilities to binary predictions as Equation 13

$$y_{i \text{ binary}}^* = \mathcal{I}(y_i^* > 0.5) \quad (13)$$

where $\mathcal{I}$ is the indicator function.

- 8: **end for**
- 9: Compute average metrics across all folds.
- 10: Output the cross-validation results.

Ensure: Cross-validation metrics: Accuracy, Precision, Recall, Specificity, Sensitivity, F1 score

termed as the marginal contribution of feature i to $\mathbb{S}$. Applying SHAP the importance of individual feature contribution can be quantified in the prospective of predicting OC. This insight enhances the openness and support evidence based clinical choices for OC detection and treatment.

4. Result Analysis

This section depicts the results of the proposed GAPO-Enhanced LightGBM model compared with eight ML models with GAPO based feature selection, such as GAPO-DT, GAPO-AdaBoost, GAPO-XGBoost, GAPO-Extra Trees RF, GAPO-Linear SVM, GAPO-Quadratic SVM, GAPO-Cubic SVM, and GAPO-LDA. Further, a comparison with the pre-existing models to ensure a comprehensive evaluation is also given. The results show the superior predictive accuracy and robustness of GAPO-Enhanced LightGBM.

4.1. Feature Selection

Feature selection plays an essential role in strengthening model efficiency, reducing overfitting, simplifying interpretation, and improving operational efficiency in the prediction of OC. Eliminating unnecessary and non-informative features, the focus is on the most insightful biomarkers. Several feature selection methods, comprising MI, LASSO, Recursive Feature Elimination (RFE), Boruta, and GAPO, are applied to the data set to extract the most pertinent features for classification. GAPO, inspired by the concept of survival of the fittest, iteratively improves the selection of characteristics by drawing the fitness of each biomarker. It emphasizes both the limitations of clinical diagnosis and its predictive value. The partial mapping of the crossover and switched mutation approaches in GAPO facilitates the selection of the best biomarkers, enhancing the detection of OC. The fitness scores for CA125, HE4, CEA, and NEU were shown to be significantly higher compared to other characteristics. GAPO achieved an optimal feature set for ten selected biomarkers, outperforming other approaches.

The GAPO has identified ten key characteristics: HE4, CA125, CEA, NEU, AFP, AST, LYM#, MPV, GLO and Menopause as the most informative for the screening of OC. Menopause is a demographic feature associated with a higher likelihood of OC, especially in postmenopausal women discussed by Pike *et al* [43]. CA125 and HE4 are well-established biomarkers for OC, with higher levels suggesting the possibility of malignant tumors, while CEA and CA19-9 distinguish between distinct tumor subtypes. Immune system indicators, such as NEU and LYM% , reflect the inflammatory response of the body, which is a prevalent feature in OC presented in [28]. Albumin, commonly low in cancer patients, serves as a measure of nutritional status and tumor burden, while GGT levels are connected with liver function and oxidative stress in metastatic cancer. Lastly, mean corpuscular hemoglobin (MCH) helps to identify anemia, which is often observed in cancer patients [29]. Together, these features provide a comprehensive profile that significantly improves the precision of OC, supporting early diagnosis and

personalized treatment planning. Table 2 presents a relative comparison of various feature selection techniques, where MI selected 32 features, Boruta selected 23, RFE chose 20 and LASSO identified 14. GAPO emerged as the optimal method to balance the significance of features, accessibility, and diagnostic efficiency. The selection process ensures the correlation of biomarkers, diagnostic reliability, test costs, and clinical restrictions, ensuring the most effective set of characteristics for the prediction of OC.

Table 2: Comparison of Feature Selection Methods: Evaluating Filtering, Wrapper, and Embedded Techniques with GAPO

Method	Class Variable	Features Selected	#Features
Mutual Information(MI)	TYPE	HE4, NEU, CA-125, Age, Menopause, ALB, LYM%, ALP, CEA, Na, MCH, HGB, CO2CP, PDW, AFP, CA19-9, BASO#, GLO, CL, PCT, Ca, MANO#, MONO%, GLU, CA72-4, PLT, TBIL, K, HCT, TP, AST, UN	32
Boruta	TYPE	AFP, Menopause, MPV, MCH, LYM#, LYM%, Na, IBIL, HE4, TP, GLO, NEU, PCT, CEA, PDW, PLT, Age, CA19-9, CA125, ALB, ALP, TBIL, AST	23
RFE	TYPE	HE4, CA125, CEA, MCH, MCV, AFP, Menopause, CA19-9, GLO, NEU, CA72-4, BASO%, BASO#, ALB, CL, CO2CP, AG, PLT, PCT, Ca	20
LASSO	TYPE	MCH, Ca, IBIL, BASO#, ALB, BUN, CREA, ALT, K, GLU, PDW, HGB, PCT, MPV	14
GAPO	TYPE	HE4, CA125, CEA, NEU, AFP, AST, LYM#, MPV, GLO, Menopause	10

4.2. Model Results

The performance of various ML models was assessed based on fundamental performance metrics, including accuracy, precision, recall, specificity, and F1-score.

Table 3 summarizes the results of the implemented models. The GAPO-DT model obtained an accuracy 0.90, precision 0.90, recall 0.90, and a specificity and F1 score of 0.89 and 0.90, respectively. Similarly, GAPO-AdaBoost also demonstrated well, with an accuracy of 0.91, a precision of 0.90, recall of 0.95, specificity 0.88 and F1 score 0.91 making it a competitive alternative. Similarly, the GAPO-XGBoost model obtained an accuracy of 0.88, with precision 0.91, recall 0.89, specificity and F1 score

Table 3: Comparison of Predictive Model’s Performance Across Different ML Approaches

Model Name	Accuracy	Precision	Recall	Specificity	F1 score
GAPO-DT	0.90	0.90	0.90	0.89	0.90
GAPO-AdaBoost	0.91	0.90	0.95	0.88	0.91
GAPO-XGBoost	0.88	0.91	0.89	0.85	0.90
GAPO-Extra Tree RF	0.91	0.90	0.94	0.88	0.91
GAPO-Linear SVM	0.85	0.80	0.96	0.73	0.87
GAPO-Quadratic SVM	0.83	0.77	0.98	0.67	0.86
GAPO-Cubic SVM	0.84	0.77	0.99	0.67	0.87
GAPO-LDA	0.83	0.76	0.99	0.65	0.82
GAPO-Enhanced LightGBM	0.92	0.90	0.96	0.88	0.93

at 0.85 and 0.90, respectively. This suggests that the GAPO-DT model is reliable but marginally less efficient compared to GAPO-AdaBoost.

The GAPO-Extra Tree RF model, on the other hand, performed marginally better with an accuracy of 0.91, a precision of 0.90, recall 0.94, specificity 0.88 and a F1 score of 0.91, suggesting potential limitations that it might not be capable of accurately detecting all positive case. While the GAPO-SVM models and GAPO-LDA demonstrate remarkably high recall, reaching up to 0.99 in some cases, their specificity score falls between 0.65 and 0.73, indicating a considerable chance of missing actual cancer cases. Specially, GAPO-linear SVM exhibited a balanced performance with accuracy 0.85, precision 0.80, recall 0.96, specificity 0.73 and F1 score 0.87 shows an adequate balance between precision and recall, while the GAPO-Quadratic SVM and GAPO-Cubic SVM, although having high recall 0.98 and 0.99, respectively, exhibits lower specificity 0.67, making them less dependable for detecting all positive cases. GAPO-LDA, with the smallest specificity 0.65, risks of overlooking 35% of real cancer cases, which could be problematic in medical diagnosis where early detection is essential. The results indicate that GAPO-Enhanced LightGBM performs better than other models, achieving an optimal accuracy 0.92, precision 0.90, recall 0.96, specificity 0.88 and F1 score 0.93. It indicates that model demonstrates well predictive capabilities with a well-balanced range of parameters across different metrics. GAPO-Enhanced LightGBM, on the other hand, significantly improves recall while retaining high precision, assuring fewer false negatives and a more consistent classification system. GAPO-Enhanced

LightGBM emerges as the most effective model, with the best trade-off between accuracy, precision, recall, and specificity. GAPO-AdaBoost and GAPO-XGBoost also provide competitive performance as well, making them viable substitutes. Meanwhile, models like GAPO-Extra Tree RF need further fine-tuning to enhance precision without preserving recall. Given the fatal consequences of misdiagnosis, GAPO-Enhanced LightGBM emerges as the most appropriate model for OC prediction.

4.3. Feature Importance

SHAP summary plot gives some important insights regarding the impact of important biomarkers involved in the diagnosis of OC as seen in Figure 4. Every dot represents a SHAP value for each feature. Features are sorted in terms of their importance with the most important ones having the most influence on the model. Feature values are represented by colors (blue=low, red=high). The positive SHAP values show that there is an increase in the model's output and negative SHAP values mean the opposite. Also, HE4, CA125, CEA, and AFP are the main tumor markers, where the presence of high levels of these markers is a clear sign of ovarian malignancies. HE4 and CA125 are especially important in the diagnosis of OC, since the increase in their levels is associated with OC development. CEA and AFP, although related to colorectal and liver cancers, are also important in assessing the risk of OC. NEU, LYM#, and GLO are immune-related markers, which influence the model's predictions based on elevated levels are strongly correlated with ovarian malignancies. HE4 and CA125 are especially important in the diagnosis of OC, since the increase in their levels is associated with OC development. CEA and AFP, although related to colorectal and liver cancers, are also important in assessing the risk of OC. NEU, LYM#, and GLO are immune-related markers, which influence the model's predictions. High NEU level could be an indicator of inflammation that is usually associated with cancer growth. On the other hand, high LYM# and GLO levels are indicators of immune responses to the development of a tumor. AST is one of the liver enzymes whose importance is seen in cases of OC

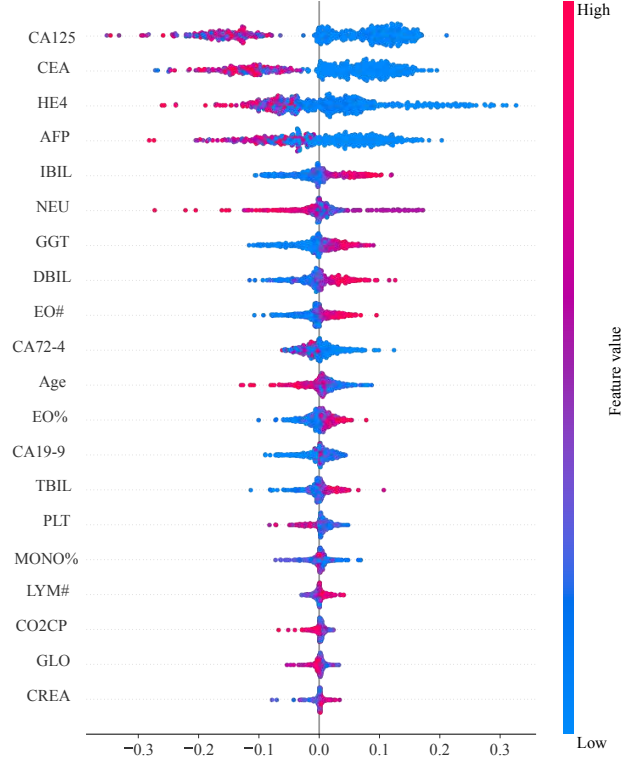


Figure 4: SHAP based visualization illustrating feature importance and impact on model predictions.

in which liver function is affected. MPV is an indicator of platelets' function. This parameter can play a significant role in recognizing any abnormality in the field of coagulation. This phenomenon occurs frequently among patients with cancer (refer Table 4).

4.4. Uncertainty Analysis

This study mainly focuses the uncertainty associated to the selection of input features in OC prediction using Enhanced LightGBM, while acknowledging additional sources of uncertainty such as performance approximation errors, sample variability, and data preprocessing biases [2]. Enhanced LightGBM, a gradient boosting approach, is well suited for considering sensitive medical data and generating both deterministic and probabilistic predictions. The uncertainty in the prediction process is reviewed through

Table 4: Overall Feature Contribution to Model Predictions

Feature	Role in Disease Prediction	High Values Indicate	Low Values Indicate
HE4	OC marker	Cancer risk	No malignancy
CA125	OC/inflammation	Cancer or inflammation	Normal levels
CEA	General tumor marker	Cancer risk	No malignancy
NEU	Immune response/inflammation	Infection or cancer	Normal immune function
AFP	Liver/testicular cancer marker	Cancer presence	No malignancy
AST	Liver/heart damage	Liver disease	Healthy liver
LYM#	Immune function	Chronic infection/inflammation	Immunosuppression
MPV	Blood clotting indicator	Risk of clotting disorders	Bleeding risk
GLO	Immune protein levels	Chronic disease	Nutritional deficiency

the change in input features and their impact on the Area Under the Curve (AUC) values. To further examine uncertainty, AUC calculation is utilized, generating multiple instances of feature perturbations and calculating the variation in AUC scores [51]. The variation in AUC is computed using Equation 15:

$$\Delta AUC = AUC_{base} - AUC_{perturbed} \quad (15)$$

where AUC_{base} is the AUC obtained using all features and $AUC_{perturbed}$ is the AUC obtained following corresponding feature perturbation. The findings illustrate that the model’s performance deteriorates considerably with feature reduction, with AUC dropping from 0.98 to 0.74. The higher value, ΔAUC , more sensitive the model is to feature selection, indicating more uncertainty. The variance of AUC values throughout these feature perturbations determines a measure of the model persistence. By incorporating these uncertainty measurement methodologies and aligning the evaluation with the observed results, the study ensures a holistic evaluation of the predictive utility of the GAPO-Enhanced LightGBM model as part of OC prediction.

Comprehending the importance of results within the uncertainty domain is fundamental, particularly when working with statistical models that yield estimates. In this study, uncertainty analysis is achieved by assessing variations in feature choices and their effects on model productivity. The Receiver Operating Characteristic (ROC) analysis highlights the influences of feature selection on predictive accuracy and uncertainties in the prediction of OC. The comprehensive assessment of how feature

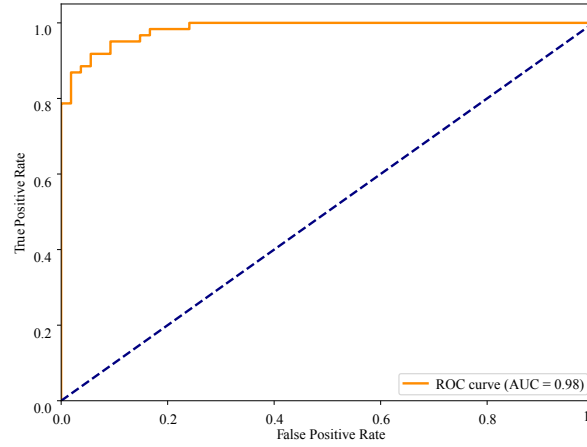


Figure 5: Original ROC illustrating the prediction of the GAPO-Enhanced LightGBM model with 10 features

reduction impacts model performance and reliability reveals significant insights into the methodical evaluation. The enhanced LightGBM procedure is adopted to evaluate uncertainty, in which different sets of features are analyzed to determine their influence on predictive stability. The use of all 10 characteristics obtained from Algorithm 1 demonstrated a high AUC of 0.98, indicating high discriminative ability as shown in Figure 5.

First, the set of characteristics was reduced to 9 characteristics by excluding the most significant biomarker CA125. The model's performance remains unchanged as shown in Figure 6. Although CA125 is a crucial biomarker for OC prediction, other characteristics share a similar discriminative power which can compensate for the missing CA125. Since AUC remains unchanged at 0.98 ± 0.01 , it suggests that the proposed model can rely on other biomarkers like HE4, AFP and CEA as they carry similar predictive information.

Reducing the set of characteristics to 8 by excluding (HE4 and CA125) other features held constant obtained in an AUC of 0.96 ± 0.01 , indicating a slight decrease in performance while preserving high precision with low uncertainty as shown in Figure 7. Further lowering the feature set to 7 by excluding (HE4, CA125 and AFP) yielded an

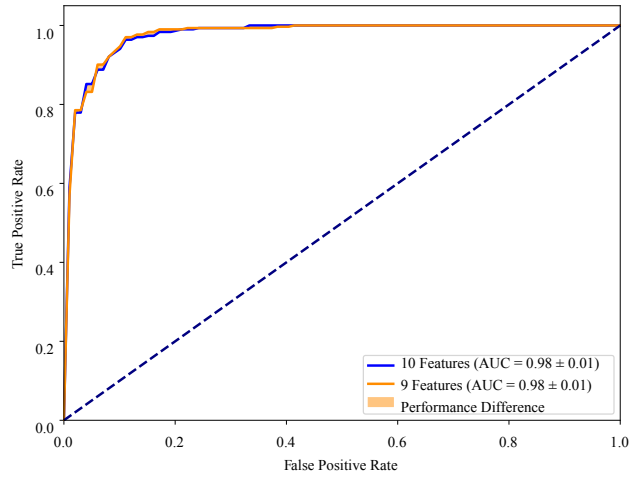


Figure 6: ROC presenting the predictive performance using reduced set of 9 features

AUC of 0.93 ± 0.01 , with a slight drop in predictive performance while still maintaining model stability shown in Figure 8.

However, with only 6 features by excluding (HE4, CA125, AFP and CEA), the AUC plummeted to 0.83 ± 0.03 , marking a discernible loss of performance and increased uncertainty, as reflected in the higher standard deviation shown in Figure 9. Lastly, the set of characteristics was reduced to 5 features by excluding (HE4, CA125, AFP, CEA and AST), the AUC remains same to 0.83 ± 0.03 , demonstrating a significant loss of important information and the highest level of uncertainty observed as shown in Figure 10.

Predictive models for OC need certain hypothesis and simplifications due to the complex nature of clinical and biological data. Analysis of uncertainties becomes necessary in order to find out and understand the limitations of such models. Uncertainty analysis provides useful results, showing that the removal of critical features leads to a drop of the AUC. The importance of carrying out uncertainty analysis for obtaining correct diagnostic results is clearly seen from this. The clinicians and decision-makers need precise and correct information to be able to make a right decision about further

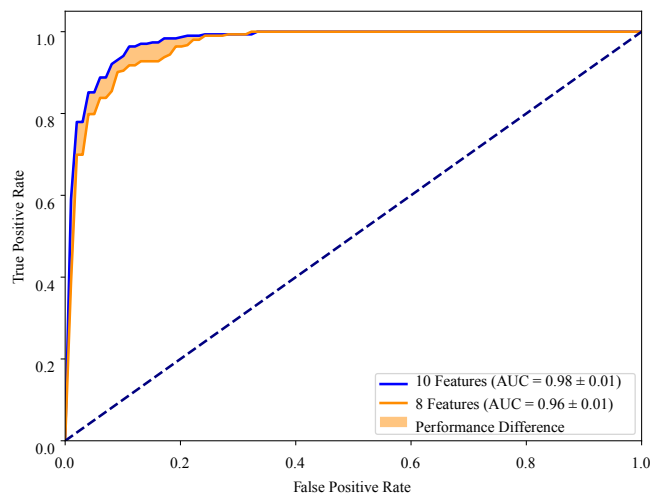


Figure 7: ROC presenting the predictive performance using reduced set of 8 features

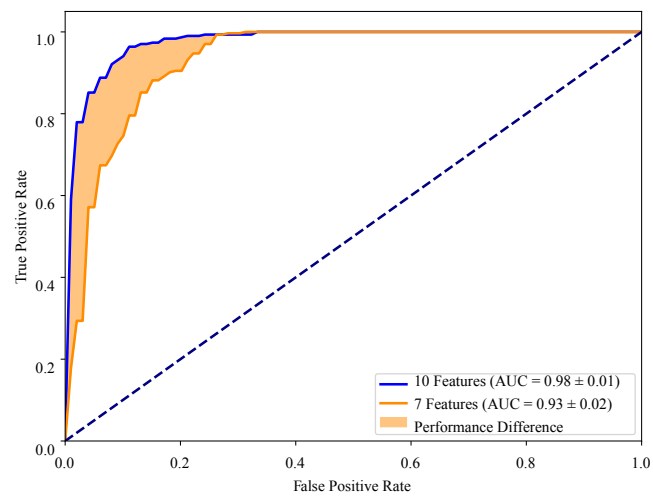


Figure 8: ROC presenting the predictive performance using reduced set of 7 features

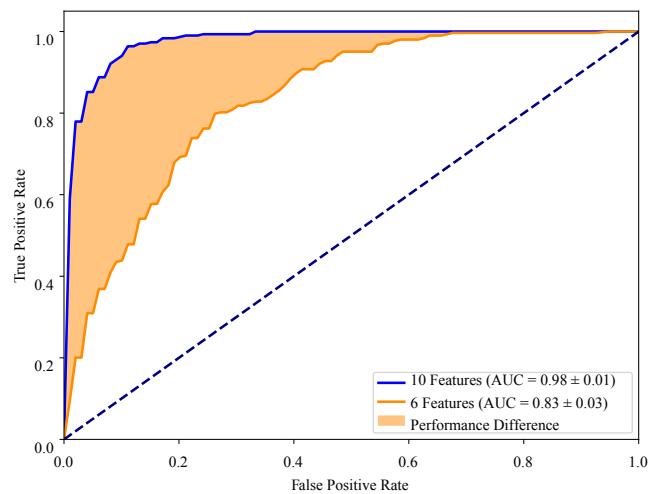


Figure 9: ROC presenting the predictive performance using reduced set of 6 features

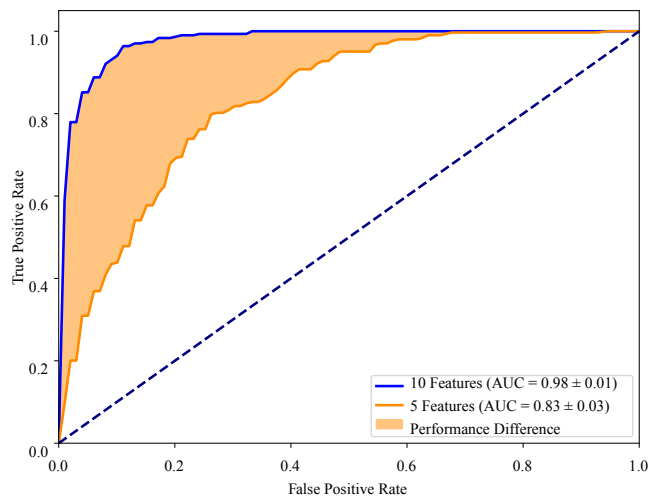


Figure 10: ROC presenting the predictive performance using reduced set of 5 features

Table 5: Comparison with existing work

Article	Data Type	Model Name	Accuracy	Precision	Recall	Specificity	F1 score
Alganmi <i>et al</i> [4]	WES Data	Random Forest	0.88	0.95	0.61	-	0.71
Akazawa <i>et al</i> [3]	Clinical Data	Extreme Gradient Boosting	0.88	0.91	0.89	0.85	0.90
Oznacaret <i>et al</i> [41]	Clinical Data	CatBoosting	0.89	0.90	0.89	-	0.89
Proposed Work	Clinical Data	GAPO-Enhanced LightGBM	0.92	0.90	0.96	0.88	0.93

treatment. Awareness of predictive uncertainty improves trust to decisions made according to model results. Moreover, clinical and genetic differences among patients may affect prediction results with the passage of time. A comparison with [4], [3] and [41] is given in Table 5.

5. Conclusion

For this research, GAPO-Enhanced LightGBM was applied to improve OC prediction results. The GAPO algorithm was applied for selecting the most useful biomarkers: HE4, CA125, CEA, NEU, AFP, AST, LYM#, MPV, GLO, and Menopause. Feature selection and classifier parameter optimization can be optimized using GAPO-Enhanced LightGBM, where the evolutionary learning can adjust the significance of features and the parameters of the classification model. Generalization approach is applied to illustrate the impact of feature selection through the uncertainty analysis, which shows the reduction of AUC from 0.98 to 0.83 after removing some important biomarkers.

This framework has been proven to greatly enhance the accuracy of the model, reaching the results of 0.92 for accuracy, 0.90 for precision, 0.96 for recall, 0.88 for specificity and 0.93 for F1 score and thus surpassing standard diagnostic models. Moreover, the robustness of the model is tested through the evaluation of the influence of different optimization methods on the accuracy of classification. As demonstrated, feature selection is a crucial element for improving the model’s predictability and transparency; here the role of SHAP in evaluating the importance of each feature is particularly crucial. It is desirable to use optimal feature selection methods, as the elimination of redundant features simplifies the classification algorithm. In addition,

in this study, selected characteristics include menopause, indicating that the risk of OC increases with age. However, our model has detected cases of OC occurring before the typical age associated with menopause. It suggests the need to refine our methods and explore additional features or approaches to enhance the effectiveness of the model in identifying OC across a broader age spectrum. Prioritizing this will enable us to develop a more robust and impact detection system.

Compliance with Ethical Standards

Ethics approval

Not Applicable

Consent to participate

Not Applicable

Research involving Human Participants and/or Animals

No

Consent for publication

Not Applicable

Funding

The authors declare that no funds, grants, or other support were received during the study.

Declaration of Interest Statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data and Code Availability

The dataset used in this study was collected from <https://data.mendeley.com/datasets/th7fztbrv9/11>.

Clinical trial number

Not Applicable.

References

- [1] Abbas, S., Asif, M., Rehman, A., Alharbi, M., Khan, M.A., Elmitwally, N., 2024. Emerging research trends in artificial intelligence for cancer diagnostic systems: A comprehensive review. *Heliyon* .
- [2] Achu, A., Aju, C., Di Napoli, M., Prakash, P., Gopinath, G., Shaji, E., Chandra, V., 2023. Machine-learning based landslide susceptibility modelling with emphasis on uncertainty analysis. *Geoscience Frontiers* 14, 101657.
- [3] Akazawa, M., Hashimoto, K., 2020. Artificial intelligence in ovarian cancer diagnosis. *Anticancer research* 40, 4795–4800.
- [4] Alganmi, N., Bashanfar, A., Alotaibi, R., Banjar, H., Karim, S., Mirza, Z., Abusamra, H., Al-Attas, M., Turkistany, S., Abuzenadah, A., 2024. Uncovering hidden genetic risk factors for breast and ovarian cancers in brca-negative women: a machine learning approach in the saudi population. *PeerJ Computer Science* 10, e1942.
- [5] Alhijawi, B., Awajan, A., 2024. Genetic algorithms: Theory, genetic operators, solutions, and applications. *Evolutionary Intelligence* 17, 1245–1256.
- [6] Ashton, J.J., Young, A., Johnson, M.J., Beattie, R.M., 2023. Using machine learning to impact on long-term clinical care: principles, challenges, and practicalities. *Pediatric Research* 93, 324–333.
- [7] Azimi-Pour, M., Eskandari-Naddaf, H., Pakzad, A., 2020. Linear and non-linear svm prediction for fresh properties and compressive strength of high volume fly ash self-compacting concrete. *Construction and Building Materials* 230, 117021.
- [8] Bagheri, M., Bagheritaba, M., Alizadeh, S., Parizi, M.S., Matoufinia, P., Luo, Y., 2024. Ai-driven decision-making in healthcare information systems: a comprehensive review. *Preprints* .

- [9] Barbieri, M.C., Grisci, B.I., Dorn, M., 2024. Analysis and comparison of feature selection methods towards performance and stability. *Expert Systems with Applications* , 123667.
- [10] Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., Lang, M., 2020. Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis* 143, 106839.
- [11] Chekroud, A.M., Zotti, R.J., Shehzad, Z., Gueorguieva, R., Johnson, M.K., Trivedi, M.H., Cannon, T.D., Krystal, J.H., Corlett, P.R., 2016. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *The Lancet Psychiatry* 3, 243–250.
- [12] Chen, R.C., Dewi, C., Huang, S.W., Caraka, R.E., 2020. Selecting critical features for data classification based on machine learning methods. *Journal of Big Data* 7, 52.
- [13] Chen, Y., Zhang, X., Grekousis, G., Huang, Y., Hua, F., Pan, Z., Liu, Y., 2023. Examining the importance of built and natural environment factors in predicting self-rated health in older adults: An extreme gradient boosting (xgboost) approach. *Journal of Cleaner Production* 413, 137432.
- [14] Choudhury, A., Mondal, A., Sarkar, S., 2024. Searches for the bsm scenarios at the lhc using decision tree-based machine learning algorithms: a comparative study and review of random forest, adaboost, xgboost and lightgbm frameworks. *The European Physical Journal Special Topics* 233, 2425–2463.
- [15] Deniz, A., Kiziloz, H.E., 2019. On initial population generation in feature subset selection. *Expert Systems with Applications* 137, 11–21.
- [16] Ellis, R.J., Sander, R.M., Limon, A., 2022. Twelve key challenges in medical machine learning and solutions.

- [17] Fan, C.Y., Chang, P.C., Lin, J.J., Hsieh, J., 2011. A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. *Applied Soft Computing* 11, 632–644.
- [18] Feng, J.z., Wang, Y., Peng, J., Sun, M.w., Zeng, J., Jiang, H., 2019. Comparison between logistic regression and machine learning algorithms on survival prediction of traumatic brain injuries. *Journal of critical care* 54, 110–116.
- [19] Ghose, A., Bolina, A., Mahajan, I., Raza, S.A., Clarke, M., Pal, A., Sanchez, E., Rallis, K.S., Boussios, S., 2022. Hereditary ovarian cancer: towards a cost-effective prevention strategy. *International journal of environmental research and public health* 19, 12057.
- [20] Havrilesky, L.J., Whitehead, C.M., Rubatt, J.M., Cheek, R.L., Groelke, J., He, Q., Malinowski, D.P., Fischer, T.J., Berchuck, A., 2008. Evaluation of biomarker panels for early stage ovarian cancer detection and monitoring for disease recurrence. *Gynecologic oncology* 110, 374–382.
- [21] Jiang, J., Meng, L., 2012. The strategy of improving convergence of genetic algorithm [j]. *Indonesian Journal of Electrical Engineering and Computer Science* 10, 2063–2068.
- [22] Juwono, F.H., Wong, W.K., Pek, H.T., Sivakumar, S., Acula, D.D., 2022. Ovarian cancer detection using optimized machine learning models with adaptive differential evolution. *Biomedical Signal Processing and Control* 77, 103785.
- [23] Katoch, S., Chauhan, S.S., Kumar, V., 2021. A review on genetic algorithm: past, present, and future. *Multimedia tools and applications* 80, 8091–8126.
- [24] Kawakami, E., Tabata, J., Yanaihara, N., Ishikawa, T., Koseki, K., Iida, Y., Saito, M., Komazaki, H., Shapiro, J.S., Goto, C., et al., 2019. Application of artificial intelligence for preoperative diagnostic and prognostic prediction in epithelial

ovarian cancer based on blood biomarkers. *Clinical cancer research* 25, 3006–3015.

- [25] Khalid, S., Khalil, T., Nasreen, S., 2014. A survey of feature selection and feature extraction techniques in machine learning, in: 2014 science and information conference, IEEE. pp. 372–378.
- [26] Khan, A., Baig, A.R., 2015. Multi-objective feature subset selection using non-dominated sorting genetic algorithm. *Journal of applied research and technology* 13, 145–159.
- [27] Khan, A.A., Chaudhari, O., Chandra, R., 2024. A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Systems with Applications* 244, 122778.
- [28] Lertkhachonsuk, A.a., Buranawongtrakoon, S., Lekskul, N., Rermluk, N., Wee-Stekly, W.W., Charakorn, C., 2020. Serum ca19-9, ca-125 and cea as tumor markers for mucinous ovarian tumors. *Journal of Obstetrics and Gynaecology Research* 46, 2287–2291.
- [29] Liu, W., Zhang, F., Quan, B., Li, M., Lu, S., Li, J., Chen, R., Yin, X., 2021. The prognostic value of the albumin to gamma-glutamyltransferase ratio in patients with hepatocellular carcinoma undergoing radiofrequency ablation. *Disease Markers* 2021, 3514827.
- [30] Lu, M., Fan, Z., Xu, B., Chen, L., Zheng, X., Li, J., Znati, T., Mi, Q., Jiang, J., 2020. Using machine learning to predict ovarian cancer. *International journal of medical informatics* 141, 104195.
- [31] Mahamud, E., Fahad, N., Assaduzzaman, M., Zain, S., Goh, K.O.M., Morol, M.K., 2024. An explainable artificial intelligence model for multiple lung diseases

- classification from chest x-ray images using fine-tuned transfer learning. *Decision Analytics Journal* 12, 100499.
- [32] Matulonis, U.A., Sood, A.K., Fallowfield, L., Howitt, B.E., Schouli, J., Karlan, B.Y., 2016. Ovarian cancer. *Nature reviews Disease primers* 2, 1–22.
- [33] Muhammad, D., Bendeche, M., 2024. Unveiling the black box: a systematic review of explainable artificial intelligence in medical image analysis. *Computational and structural biotechnology journal* .
- [34] Nadirgil, O., 2023. Carbon price prediction using multiple hybrid machine learning models optimized by genetic algorithm. *Journal of Environmental Management* 342, 118061.
- [35] Nagassou, M., Mwangi, R.W., Nyarige, E., 2023. A hybrid ensemble learning approach utilizing light gradient boosting machine and category boosting model for lifestyle-based prediction of type-ii diabetes mellitus. *Journal of Data Analysis and Information Processing* 11, 480–511.
- [36] Nematzadeh, H., García-Nieto, J., Navas-Delgado, I., Aldana-Montes, J.F., 2024. Feature selection using a classification error impurity algorithm and an adaptive genetic algorithm improved with an external repository. *Knowledge-Based Systems* 301, 112345.
- [37] Noei Teymoordash, S., Zendehdel, H., Norouzi, A.R., Kashian, M., 2025. Diagnostic accuracy of artificial intelligence algorithms to predict remove all macroscopic disease and survival rate after complete surgical cytoreduction in patients with ovarian cancer: a systematic review and meta-analysis. *BMC surgery* 25, 27.
- [38] Nopour, R., 2024. Screening ovarian cancer by using risk factors: machine learning assists. *BioMedical Engineering OnLine* 23, 18.

- [39] Oliveira, A.L., Braga, P.L., Lima, R.M., Cornélio, M.L., 2010. Ga-based method for feature selection and parameters optimization for machine learning regression applied to software effort estimation. *information and Software Technology* 52, 1155–1166.
- [40] Omotehinwa, T.O., Oyewola, D.O., Mounq, E.G., 2024. Optimizing the light gradient-boosting machine algorithm for an efficient early detection of coronary heart disease. *Informatics and Health* 1, 70–81.
- [41] Öznacar, T., Güler, T., 2025. Prediction of early diagnosis in ovarian cancer patients using machine learning approaches with boruta and advanced feature selection. *Life* 15, 594.
- [42] Piedimonte, S., Mohamed, M., Rosa, G., Gerstl, B., Vicus, D., 2025. Predicting response to treatment and survival in advanced ovarian cancer using machine learning and radiomics: A systematic review. *Cancers* 17, 336.
- [43] Pike, M.C., Pearce, C.L., Peters, R., Cozen, W., Wan, P., Wu, A.H., 2004. Hormonal factors and the risk of invasive ovarian cancer: a population-based case-control study. *Fertility and sterility* 82, 186–195.
- [44] Rosenthal, A.N., Menon, U., Jacobs, I.J., 2006. Screening for ovarian cancer. *Clinical obstetrics and gynecology* 49, 433–447.
- [45] Rostami, M., Berahmand, K., Forouzandeh, S., 2021. A novel community detection based genetic algorithm for feature selection. *Journal of Big Data* 8, 2.
- [46] Seyyedattar, M., Zendeboudi, S., Ghamartale, A., Afshar, M., 2024. Advancing hydrogen storage predictions in metal-organic frameworks: a comparative study of lightgbm and random forest models with data enhancement. *International Journal of Hydrogen Energy* 69, 158–172.

- [47] Sharma, D., Kumar, R., Jain, A., 2022. Breast cancer prediction based on neural networks and extra tree classifier using feature ensemble learning. *Measurement: Sensors* 24, 100560.
- [48] Sun, Y., Wen, B., 2024. Machine-learning diagnostic models for ovarian tumors. *Heliyon* 10.
- [49] Tan, F., Fu, X., Zhang, Y., Bourgeois, A.G., 2008. A genetic algorithm-based method for feature subset selection. *Soft Computing* 12, 111–120.
- [50] Tsai, C.F., Eberle, W., Chu, C.Y., 2013. Genetic algorithms in feature and instance selection. *Knowledge-Based Systems* 39, 240–247.
- [51] Velikova, T., Mileva, N., Naseva, E., 2024. Method “monte carlo” in healthcare. *World Journal of Methodology* 14, 93930.
- [52] Whitwell, H.J., Worthington, J., Blyuss, O., Gentry-Maharaj, A., Ryan, A., Gunu, R., Kalsi, J., Menon, U., Jacobs, I., Zaikin, A., et al., 2020. Improved early detection of ovarian cancer using longitudinal multimarker models. *British journal of cancer* 122, 847–856.
- [53] Ying, C., Qi-Guang, M., Jia-Chen, L., Lin, G., 2013. Advance and prospects of adaboost algorithm. *Acta Automatica Sinica* 39, 745–758.
- [54] Zhang, Y.P., Zhang, X.Y., Cheng, Y.T., Li, B., Teng, X.Z., Zhang, J., Lam, S., Zhou, T., Ma, Z.R., Sheng, J.B., et al., 2023. Artificial intelligence-driven radiomics study in cancer: the role of feature engineering and modeling. *Military Medical Research* 10, 22.