Comparative Analysis of Machine Learning Frameworks for Robust Ovarian Cancer Detection Using Feature Selection and Data Balancing

DSS LakshmiKumari P[1], Maragathavalli P[2]
Information Technology, Puducherry Technological University, Pillaichavady, Puducherry 605014[1,2]
SRKR Engineering College, Bhimavaram[1]

***Abstract*—One of the most serious malignancies that affects women's health worldwide is ovarian cancer. As a result, prompt accurate diagnosis and treatment are necessary. This study's primary objective is to determine whether or not OC is present within the body of a person by using a range of characteristics gleaned with a couple of health examinations. The article is concentrated on twelve ML techniques used for OC diagnosis. The dataset has been altered by applying the borderline SVMSMOTE method to address the imbalance properties and the MICE imputation method to impute the missing values in order to enhance the performance of the classifiers. Additionally, the boruta approach and recursive feature reduction has been utilized to identify the most important features while the hyper parameter tuning strategy has been employed to improve classifier performance and provide ideal solutions.Boruta opted just 50% of the total characteristics and outperformed RFE while considering the most important feature. Furthermore, many performance measures are used to determine which classifiers are the best in identifying OC. Voting classifier surpassed state-of-the-art approaches and other machine learning methods with the highest accuracy. The suggested approach obtained the highest average of 93.06% accuracy, 88.57% precision, 96.88% recall, 92.54% F1-score, and 93.44% AUC-ROC based on experimental results. Experiments show that in comparison with the state-of-the-art techniques, our suggested method can identify OC more accurately.**

Keywords*—Ovarian cancer detection; machine learning framework; data balancing; feature selection

I. Introduction

Globally, ovarian cancer (OC) is one of the most common cancers in women. As to the most recent worldwide cancer data published in 2021, there are roughly 314,000 new ovarian cancer cases (3.4% of all new cancer cases in women) and 207,000 ovarian cancer deaths annually (4.7% of all women's cancer deaths) [1]. Most of the time, ovarian cancer is discovered in its later stages due to a absence of specific symptoms and indicators as well as improper screening, which frequently results in unfavorable consequences [2], [3]. In general, gynaecologists have to determine if a patient has grown malignant pelvic lumps, which may indicate a tumour. Although certain approaches such as h CT and ultrasound have been used to differentiate benign tumors from malignant non-gynecologic conditions, tumor biomarkers such as human epididymis protein 4 (HE4), carbohydrate antigen 125 (CA125) and carbohydrate antigen 72-4 (CA72-4) are critical in differentiating between pelvic masses of women [4]. To stratify patients with OC into benign and malignant categories, machine learning (ML) models such as Light Gradient Boosting Machine (LGBM), Gradient Boosting Machine (GBM), Extreme Gradient Boosting Machine (XGBoost), Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), and are utilised in the construction of classification models [5].

Compared to traditional methods, ML techniques are extremely useful for identifying complicated patterns in biomarkers yet, they may not provide physical insights into diagnosis. The use of eXplainable Artificial Intelligence (XAI) can be enhanced by gaining a greater understanding of how sophisticated machine learning algorithms make decisions [6].

The main goal of this study is to identify OC using a variety of characteristics obtained from certain clinical examinations. Consequently, two datasets are used to test ensemble learning techniques, including RF, gradient boosting decision tree (GBDT), adaptive boosting,voting, bagging, XGBoost, LightGBM, and stacking. These ensemble learning techniques weren't used in parallel to identify OC in the earlier research publications. The models' performance is assessed using the dataset that is accessible to the general audience.

Various metrics, including precision, accuracy, recall, AUCROC, and F1-score, are employed to evaluate the effectiveness with every classifier.

The primary contributions of this study are as follows:

- In order to increase the effectiveness of the existing ovarian cancer diagnosis method, this research offered eight machine learning algorithms. In order to enhance performance, we have also adjusted each classifier's hyperparameter through the parameter tuning procedure.
- This study employed the MICE imputation method to address the missing information instead of applying a statistical approach. Additionally, the Borderline-SMOTE approach was used to balance the uneven data. Moreover, the relevant features are selected using recursive feature elimination (RFE) and Boruta approaches. Such OC prediction methods have not previously been thoroughly examined.

The structure of this work is as follows: Section 2 is devoted to the literature review, and Section 3 details the materials and procedures. Comparably, Section 5 presents the experimental findings and a discussion, while Section 4 shows

TABLE I. COMPARATIVE ANALYSIS OF MACHINE LEARNING APPROACHES FOR OVARIAN CANCER DETECTION

RefID	Dataset	Features	Feature Selection	Classifier	Evaluation Metrics
[4]	Mendeley	2	MRMR	DT, ROMA, LR	Acc: 92%, 95%, 97%
[5]	Mendeley	OC markers	T-test, Mann–Whitney U-test	Random Forest	Acc: 91%
[6]	PCLO Biomarker	-	-	Random Forest	AUC-ROC: 71%
[7]	Ultrasound Images	-	Fast Fourier Transform	LD, SVM, ELM	Acc: 85%
[8]	PCLO	27	F-test	SVMSMOTE	PPV: 90%
[9]	Tokyo Medical University	16	-	XGBoost	Acc: 80%
[11]	Mendeley	5	-	Group MCP	Acc: 93.33%, AUC: 0.892
[17]	Mendeley	16	SHAP	Stacking Classifier	Accuracy: 96.87%

the experimental setup. Followed by Section 6 represents the conclusion and ongoing research, in summary.

II. LITERATURE SURVEY

This section walks around a number of cutting-edge machine learning techniques that have been created to identify ovarian cancer. These strategies cover a variety of techniques, such as deep learning methods, supervised learning, and unsupervised learning. Through the utilisation of ML, these approaches have the possibility to enhance the effectiveness, precision, and expand-ability of ovarian cancer detection and diagnosis, ultimately leading to improved patient results. Table I presents a summary of the relevant works.

The author Mart´ınez-M´as et al. [7] identified the optimal classifier for FT-based feature descriptors and compare the ELM algorithm with traditional classifiers, by using ultrasound of 348 images. They has demonstrated that machine learning techniques, including LD and SVM approaches, are effective in creating the classification phase with 86% accuracy.

The most reliable predicting method for ovarian cancer, according to the Yang et al. [8] study, was the decision tree using SVMSMOTE on PLCO imbalanced dataset of143:7 class ratio. It had PPV of 0.9041, AUC of 0.9532, sensitivity of 0.7792, and specificity of 0.9982 is achieved by considering only seven features where there is chance of overfitting.

The possibility of applying the XGBoost machine learning algorithm to predict ovarian tumor which is proposed by Akazawa and Hashimoto[9], attained an accuracy of 0.80 for 16 characteristics on the 202 patient data set. On other hand Charkhchi et al. [10]focused the role that CA125 plays in the treatment of ovarian cancer in a review article , the necessity of accessible and affordable screening methods for the quick recognition of type II tumors. To predict benign or malignant ovarian tumors, the authors Hu et al. [11] selected distinct 11 groups by using 349 patients of 46 features achieved an accuracy of 93.33on other hand Ghoniem et al. [12] proposed to predict subtypes of other cancers by combining characteristics from pathologic image modal and gene modal. TCGAOV dataset is used containing pathological pictures, copy number variations, and gene expression data for 587 OC patients.The author Sorayaie Azar et al. [13] by using the SEER dataset, they study created ML techniques for regression and classification methods. XGBoost and RF were themost successful approaches for regression and classification, respectively. The findings showed promise and could be a useful tool for physicians to better understand the status of patients with ovarian cancer. For the purpose of OC detection, the authors Juwono et al. [14] has provided the best simultaneous feature weighting/parameter optimisation method. The LASSO regularisation and adaptive differential evolution (ADE) with cross validation error are used as the fitness function to optimise the weights.On other hand Boehm et al. [15] designed Decision support system to find three highly discriminating proteins (TOP1, PDIA4, and OGN) for the diagnosis of highgrade serous ovarian cancer (HGSOC) utilizing proteome data and a machine learning-based methodology. The DSS demonstrated promising clinical utility in the diagnosis of HGSOC with its excellent specificity and sensitivity in differentiating tumor from non-tumor patients.

On other hand Comes et al. [16] implemented an explainable machine learning model, the study identified high-risk BRCA-mutated patients and established when RRSO should be performed. The model produced encouraging results, with accuracy, specificity, sensitivity, G-mean, and AUC values of 83.2%, 85.3%, 57.1%, and 71.1%, respectively.

The key findings by the Abuzinadah et al. [17] include the extremely accurate stacked ensemble model for ovarian cancer prediction, the intricate interactions between biomarkers shown by the SHAP analysis, and the potential for real-world application.

III. METHODS AND MATERIALS

This article has built a smart approach for assessing and identification of OC. This study analyses OC data in twelve different ways. AUC-ROC, F-measure, accuracy, precision, recall, and F-measure were among the metrics used to assess each model's performance. The dataset was sourced from the Mendeley repository website.

A pair feature selection technique Boruta and Recursive feature elimination are employed to determine which features are most relevant. The steps involved in this investigation are shown in Fig. 1. Assigning incomplete values to every attribute in the dataset, the MICE method and a basic imputer are used. After the target attributes are equalised using the borderline-SMOTE technique, the processed data is fed into each classifier. Each classifier's parameter is adjusted via the hyper parameter tuning technique and randomised grid search. In this research, cross-validation techniques are used to validate the model functionality.

A. Dataset Description

The Third Affiliated Hospital of Soochow University provided the dataset for the study, which included 349 patients.

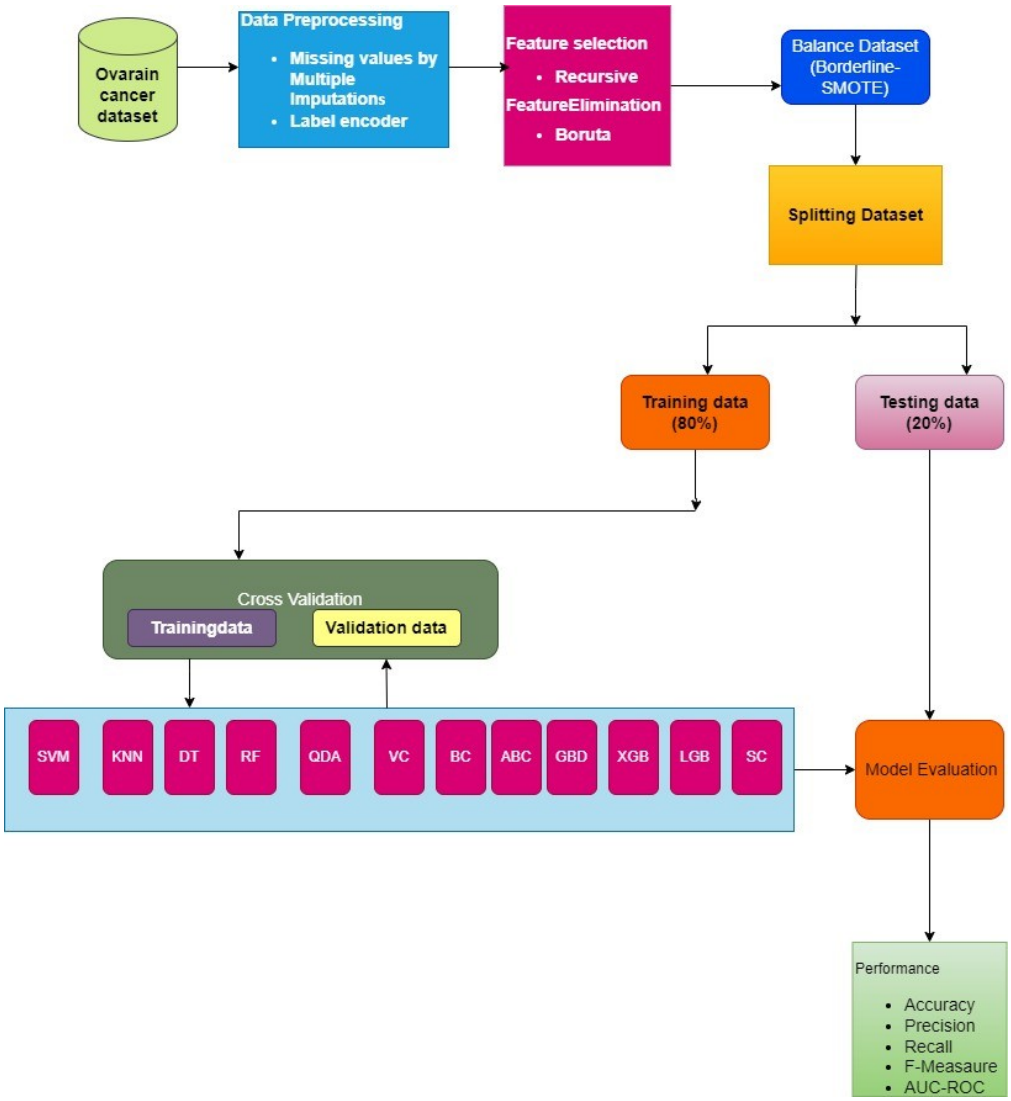

Fig. 1. Framework of detecting ovarian cancer.

The information was gathered between July 2011 and July 2018 and was split into two categories: 178 patients with benign ovarian tumour diagnoses, and 171 patients with malignant diagnoses [5].

B. Data Preprocessing

Preprocessing is necessary for the majority of real-world data because of outliers, noisy characteristics, missing values, and inconsistent patterns. If not, it will be difficult to increase the machine model's quality and produce an acceptable outcome. Numerous procedures have been used in this part to enhance and clean the data.

Several feature columns are labelled as inputs in the case of MICE technique, whereas single feature column that contains data missing is considered as an output. Subsequently, the output data is predicted using the regression model, and this is done again. Every value that is missing in the dataset is imputed using a different value in each iteration, and the process is repeated until convergence is achieved.

1) Missing value imputation: In this representation D represents the dataset with missing values, M represents the number of imputations to generate, k represents the number of iterations for each imputation, D_{imputed} represents the imputed dataset. The MICE algorithm iteratively imputes missing values for each variable in the dataset using regression models based on observed values of other variables. It repeats this procedure for a definite number of iterations for each imputation. Finally, it returns a list of M imputed datasets.

By using Multiple Imputation method [Algorithm 1], the dataset which contains 331 missing values which are been replaced.

The equation for imputing missing values using linear regression can be represented as follows:

$$\hat{Y}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} \quad (1)$$

where:

- $\hat{Y}_i$ — predicted value for the i^{th} observation.
- β_0 — intercept term.
- $\beta_1, \beta_2, \ldots, \beta_p$ — regression coefficients.
- $X_{i1}, X_{i2}, \ldots, X_{ip}$ — observed values of p predictors for observation i.

Algorithm 1 Multiple Imputations by Chained Equations (MICE)

procedure MULTIPLEIMPUTATIONSBYCHAINEDEQUATIONS(D, M, k)
  Input:
    D: Dataset with missing values
    M: Number of imputations to generate
    k: Number of iterations for each imputation
  Output: List of M imputed datasets
  Initialize empty list: $imputed\_datasets \leftarrow [\,]$
  for $i = 1$ **to** M **do**
    Initialize $D_{\text{imputed}} \leftarrow D$ ▷ Copy of original dataset
    for $j = 1$ **to** k **do**
      for all variables v with missing values in D_{imputed} **do**
        Predict missing values in v using linear regression
        Update v in D_{imputed} with predicted values
      end for
    end for
    Append D_{imputed} to $imputed\_datasets$
  end for
  return $imputed\_datasets$
end procedure

C. Feature Selection

The process of eliminating unnecessary characteristics from a machine in order to reduce its computational load and improve the efficiency of the machine model is known as feature selection. The dual feature selection methods are used in this research.

1) Recursive Feature Elimination (RFE): This method selects the best subset of features from the entire dataset and minimizes the number of features. The SelectOptimalFeatures algorithm [Algorithm 2] provides a detailed explanation of each step in the RFE method.

2) Boruta based on RF: The BorutaFeatureSelection algorithm iteratively creates shadow features, combines them with original features, fits a Random Forest model, computes Z scores, and marks features as important or unimportant based on the comparison of Z scores. It repeats this procedure for a specified number of iterations and returns the desired number of optimal feature subsets.

3) Boruta based on RF: The BorutaFeatureSelection algorithm [Algorithm 3] iteratively creates shadow features, combines them with original features, fits a Random Forest model, computes Z scores, and marks features as important or unimportant based on the comparison of Z scores. It repeats this procedure for a specified number of iterations and returns the desired number of optimal feature subsets.

Algorithm 2 Select Optimal Features

Require: D: Training dataset with m samples and r features
  Y: Output values
  F: Set of features
  k: Number of iterations
  n: Number of optimal feature subsets
Ensure: F_{prime}: n optimal feature subsets
  Procedure:
  1. Train a RF model on dataset D.
  2. Initialize the feature subset F_{prime} with all features ($F_{\text{prime}} = F$).
  for $i = 1$ to k **do**
    a. Compute coefficients using Logistic Regression (LR) on the feature subset F_{prime}.
    b. Calculate the rank R based on the coefficients.
    c. Identify a specific number of features to eliminate, such as those with the smallest coefficients.
    d. Remove the determined number of features with the smallest coefficients from F_{prime}.
  end for

Algorithm 3 Boruta Feature Selection

procedure BORUTAFEATURESELECTION(D, Y, F, k, n)
  Input:
    D: Training dataset with m samples and r features
    Y: Target values
    F: Set of features
    k: Number of iterations
    n: Number of optimal feature subsets to select
  Output:
    F': Top n important features
  Initialize $selected\_features \leftarrow []$
  for $i = 1$ **to** k **do**
    Generate shadow features by shuffling columns of F
    Combine F and shadow features to form dataset X
    Train Random Forest (RF) model on X
    Compute Z-scores: Z_o (original features), Z_s (shadow features)
    Let Z_{max} be the maximum of Z_s
    for each feature f_i in F **do**
      if $Z_o[f_i] > Z_{\text{max}}$ **then**
        Mark f_i as important
      else
        Mark f_i as unimportant
      end if
    end for
    Append important features to $selected\_features$
  end for
  Select top n features from $selected\_features$
  return F'
end procedure

IV. Classification Models

In this investigation, a variety of supervised ensemble-based machine learning algorithms [19] have been utilized, such as LR, DT, RF, LGBM, SVM, XGB, and GBM, in isolation, to forecast ovarian cancer. Furthermore, the search was conducted to identify the models that demonstrated the most outstanding performance.

1) SVMs: Supervised learning algorithms that are applied to regression and classification problems. It determines the best hyperplane with the largest margin between data points belonging to distinct classes. Using various kernel functions, SVM can handle both linear and nonlinear classification tasks.

2) k-NN: A feature space classification algorithm that classifies data points by using the majority class of the k nearest neighbors. This algorithm is regarded as non-parametric as it doesn't rely on explicit models.

3) Tree: This is probably a reference to decision trees, which divide the feature space into regions recursively and give each region a class name. Although decision trees are simple to understand, overfitting is a possibility.

4) RF: A technique that constructs many decision trees and combines their predictions by averaging or voting. This method successfully reduces overfitting by adding randomness to the feature and data sample selection processes.

5) QDA: A classification process known as quadratic discriminant analysis, or QDA, uses a quadratic decision boundary to simulate the probability density function of each class. It functions on the presumption that each class's features follow a normal distribution.

6) Voting: An ensemble learning technique where numerous models are trained independently and then their forecasts are combined by either average (for regression) or majority voting (for classification).

7) Bagging: Bagging (Bootstrap Aggregating) is an ensemble learning technique that uses bootstrap sampling of the training data to generate many models (typically of the same kind). To increase generalization performance and minimize variance, the predictions of the models are combined by voting or averaging.

8) AdaBoost: A technique that turns a number of ineffective learners into one powerful one. In each iteration, it gives more weight to data points that were incorrectly classified as it trains models progressively.

9) GBMs: Gradient Boosting Machines, or GBMs, are an ensemble learning technique that constructs a series of weak learners, usually decision trees, one after the other in a staged fashion, with each new tree fixing the mistakes of the preceding one.

10) XGBoost: A meticulously optimized implementation of the gradient boosting technique, specifically designed to enhance both speed and performance. By incorporating a diverse range of regularization techniques, XGBoost effectively prevents overfitting, thereby enabling it to achieve state-of-the-art outcomes in numerous machine learning competitions.

11) LGB: An additional scalable and extremely effective gradient boosting framework. To minimize memory usage and expedite training, it employs histogram-based approaches and a unique tree-growing algorithm.

12) SC: Also known as stacked generalization or stacking ensemble, is a machine learning technique that combines multiple classification models to improve prediction performance. Instead of relying on just one model, it leverages the collective wisdom of several models to make predictions.

V. Results

The recommended models are run through the Python program on a Windows 10 computer for validation. The system is powered by an Intel Core i3 processor with a 2.40GHz clock frequency and 8GB of RAM.

Additionally, this section involves in splitting the data source into 20% test and 80% training sets, adjusting the parameters of each classifier using the randomized grid search technique, and evaluating metrics to verify the performance of the model.

A. Parameter Tuning

Hyper-parameter optimization, or parameter tuning, is an essential stage in the creation of machine learning models. In order to maximize an algorithm's performance on a particular dataset, the optimal set of hyper-parameters for that method must be chosen. Additionally, the optimal subset of hyper-parameters for every classifier was shown in Table II to yield the best results for the OC prediction.

B. Model Evaluation Metric

The accuracy of a model's predictions can be assessed using a variety of performance evaluation criteria. More significantly, the well-known used performance evaluation metrics for binary classification include confusion matrix, precision, AUC-ROC, recall, F-measure, and accuracy.

1) Confusion matrix: The effectiveness of a classification algorithm is summarized in a table known as a confusion matrix that counts of false positives (NP), false negatives(NN), true positives (PP), and true negatives(PN).

2) Accuracy: It is the measure the proportion of a given class that is correctly classified among all instances is given by,

$$Accuracy = \frac{PP + NN}{PP + NN + PF + NF} \tag{2}$$

3) Recall: It is the proportion of all positively predicted instances to accurately predicted positive cases is given by,

$$Recall = \frac{PP}{PP + NN} \tag{3}$$

4) Precision: It is the ratio of correctly predicted positive cases among all positive cases is given by,

$$Precision = \frac{PP}{PP + PN} \tag{4}$$

TABLE II. LIST OF PRIMARY HYPERPARAMETERS FOR BINARY CLASSIFICATION FOR EACH CLASSIFIER

Classifier	Hyperparameters
SVM	kernel='rbf', C=10, gamma=0.1, probability=True
KNN	n_neighbors=11
RF	n_estimators=60, max_features=None, bootstrap=True, min_samples_split=3, min_samples_leaf=3, max_depth=None, random_state=0
QDA	Default parameters
DT	criterion='gini', min_samples_split=4, random_state=0, max_depth=None
SC	estimators=[('svc', svm), ('knn', knn)], final_estimator=rf
VC	estimators=[('svc', svm), ('knn', knn), ('rf', rf)], voting='soft'
BC	n_estimators=500, base_estimator=tree, n_jobs=-1, bootstrap=True, random_state=0, bootstrap_features=False
ABC	base_estimator=tree1, n_estimators=500, algorithm='SAMME', random_state=0
GDB	learning_rate=0.1, n_estimators=300, max_depth=3, min_samples_leaf=10
LGB	learning_rate=0.1, n_estimators=500, max_depth=3, min_child_weight=3
XGB	learning_rate=0.1, n_estimators=200, max_depth=2, min_child_weight=3

TABLE III. PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS USING ALL FEATURES

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	ROC AUC (%)
SVM	79.1	79.3	71.9	75.4	78.4
KNN	77.8	68.2	93.8	78.9	79.4
DT	83.3	79.4	84.4	81.8	83.4
RF	87.5	81.1	93.8	87.0	88.1
QDA	86.1	82.4	87.5	84.8	86.2
VC	90.3	87.9	90.6	89.2	90.3
BC	88.9	83.3	93.8	88.2	89.4
ABC	83.3	85.7	75.0	80.0	82.5
GDB	87.6	82.9	90.6	86.6	87.8
XGB	88.9	85.3	90.6	87.9	89.1
LGB	90.3	87.9	90.6	89.2	90.3
SC	76.4	74.2	71.9	73.0	75.9

5) F1-score: It is harmonic mean of precision and recall is given by,

$$F1 - score = \frac{1}{\text{Precision}} + \frac{1}{\text{Recall}} \tag{5}$$

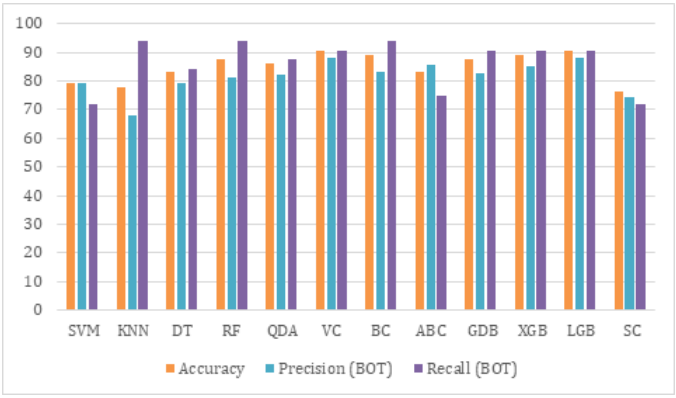

Fig. 2. Comparison of accuracy, Precision and recall for ML models of all features.

C. Results of Without Feature Selection Method

The experimental outcomes of each classification model employing every feature in the dataset are shown in Table III, Fig. 2 and Fig. 3. It was evident that the Voting classifier and LightGBM achieved a 90.3% score and all the other performance metrics are high except recall. Moreover, the other classifiers Bagging classifier, XGBoost, GBM, RF are performing better and their scores are approximately 87.5%.

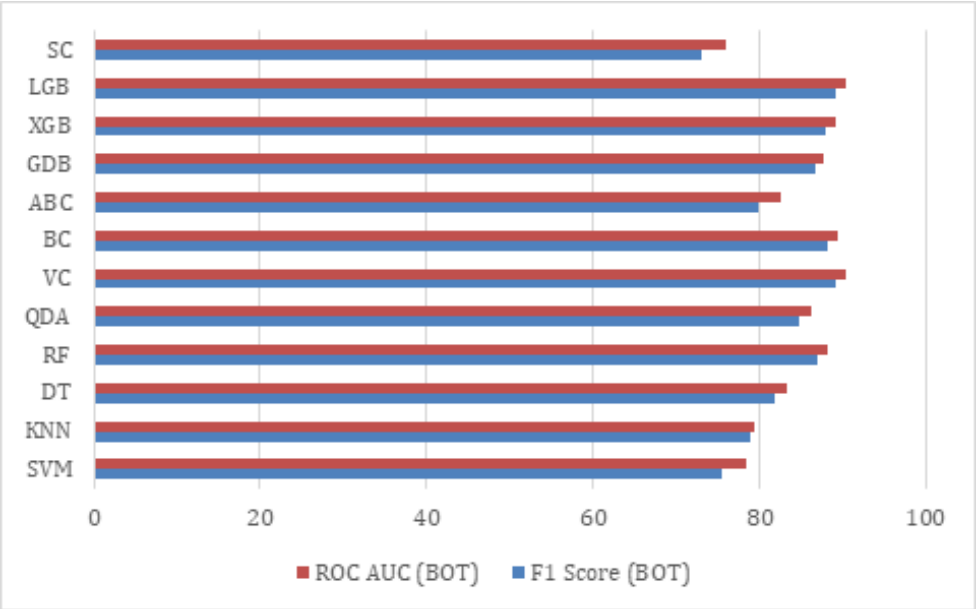

Fig. 3. Comparison of ROC and F1-Score for ML models of all features.

TABLE IV. PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS USING RECURSIVE FEATURE ELIMINATION (RFE)

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	ROC AUC (%)
SVM	80.6	78.1	78.1	78.1	80.3
KNN	76.4	71.4	78.1	74.6	76.6
DT	75.0	71.9	71.9	71.9	74.7
RF	88.9	83.3	93.8	88.2	89.4
QDA	70.8	76.2	50.0	60.4	68.8
VC	**93.1**	88.6	**96.9**	**92.5**	**93.4**
BC	86.1	80.6	90.6	85.3	86.6
ABC	87.5	84.8	87.5	86.2	87.5
GDB	86.1	80.6	90.6	85.3	86.6
XGB	87.5	84.8	87.5	86.2	87.5
LGB	87.5	84.8	87.5	86.2	87.5
SC	72.2	68.8	68.8	68.8	71.9

D. Results of Recursive Feature Elimination Method

Among all the features in the present investigation, twenty-four are significant features. AFP, Age, ALB, ALP, AST, CA125, CA19-9, CEA, CO2CP, GLO, HE4, HGB, IBIL, LYM#, LYM%, MCH, Menopause, MPV, Na, PCT, PHOS, PLT, TBIL, TP are considered for classification of OC. It was observed from Table IV that Voting classifier achieved 93.1% accuracy, 88.6% precision, 96.9% recall, 92.5% F1 score, 93.4% ROC AUC among all the ensemble classifiers. Moreover, RF, XGB, LGB have moderate scores. Based on rank important features are identified using RFE are shown in Fig. 4.

E. Results of Boruta Feature Selection Method

In this feature selection method of all features, twenty-one are important features (AFP, Age, ALB, ALP, AST, CA125, CA19-9, CEA, GLO, HE4, IBIL, LYM#, LYM%, MCH, Menopause, MPV, Na, PCT, PLT, TBIL, TP) are considered for classification of OC. It was observed from Table V that the Voting classifier achieved 93.06% accuracy, 88.57% precision, 96.88% recall, 92.54% F1 Score, 93.44% ROC AUC among all the ensemble classifiers. Moreover, RF, AdaBoost, XGBoost have moderate scores approximately 91%. Based on rank important features are identified using Boruta are shown in Fig. 5; moreover, the common features among both approaches are shown in Fig. 6.

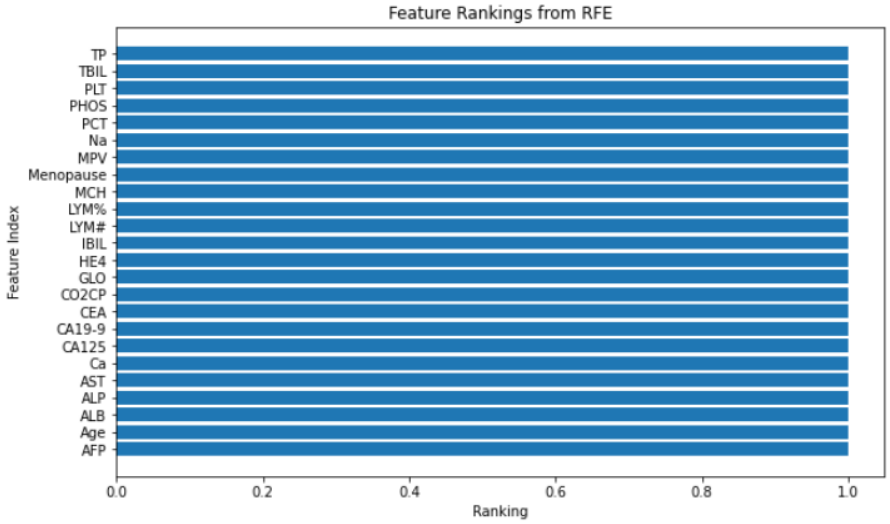

Fig. 4. Important features selected using Recursive Feature Elimination (RFE).

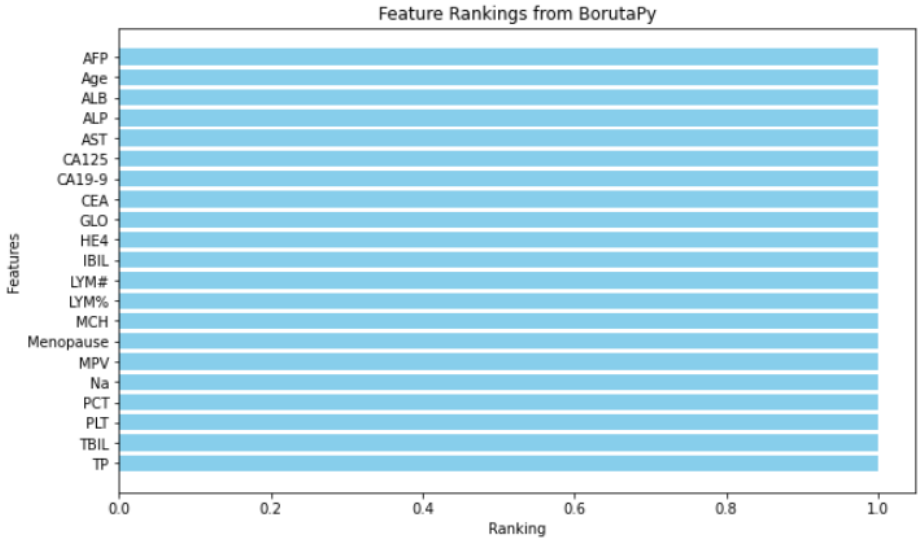

Fig. 5. Important features identified using Boruta.

TABLE V. PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS USING BORUTA FEATURE SELECTION METHOD

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	ROC AUC (%)
SVM	79.17	77.42	75.00	76.19	78.75
KNN	73.61	66.67	81.25	73.24	74.38
DT	80.56	75.00	84.38	79.41	80.94
RF	91.67	88.24	93.75	90.91	91.88
QDA	88.89	92.86	81.25	86.67	88.13
VC	**93.06**	88.57	**96.88**	**92.54**	**93.44**
BC	88.89	85.29	90.62	87.88	89.06
ABC	90.28	87.88	90.62	89.23	90.31
GDB	88.89	85.29	90.62	87.88	89.06
XGB	90.28	**90.32**	87.50	88.89	90.00
LGB	86.11	84.38	84.38	84.38	85.94
SC	81.94	88.00	68.75	77.19	80.63

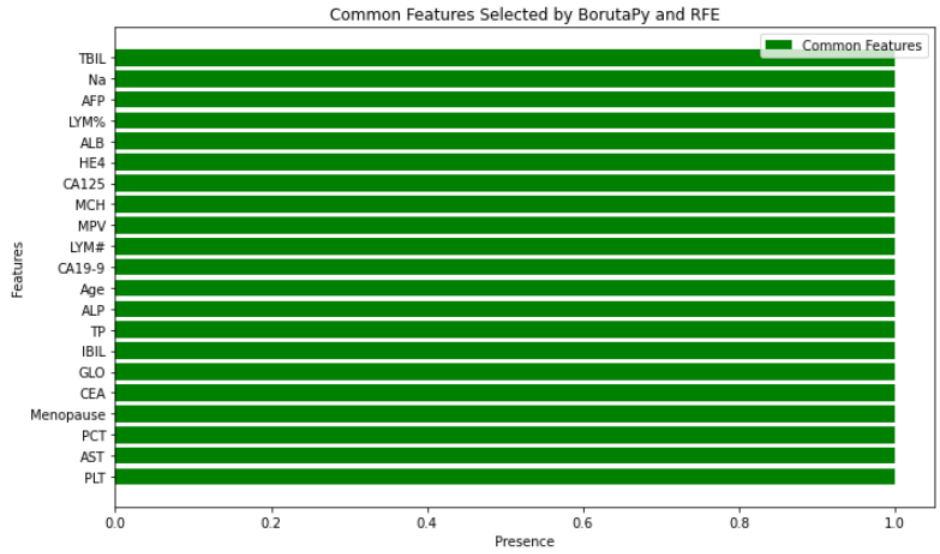

Fig. 6. Common features between Boruta and RFE.

TABLE VI. COMPARISON OF THE PROPOSED MODEL WITH EXISTING SIMILAR WORKS

Ref.ID	Missing Values	Feature Selection	Model	Precision	Recall	F1-Score	ROC-AUC	Accuracy
[17]	Mean	SHAP	GBM+XGB	88.68	88.49	88.69	–	88.24
[5]	Mean	t-test,u-test	RF	83	95	89	87	88
[11]	Mean	11 groups	Group MCP	84.6	88	–	–	93
[18]	Mean	SHAP	LGBM	89	94	91	95	91
Our Method-1	MICE	–	LightGBM	87.9	90.6	89.2	90.3	90.3
Our Method-2	MICE	RFE	VC	88.6	96.9	92.5	93.4	93.01
Our Method-3	MICE	Boruta	VC	88.57	96.88	92.54	93.44	93.06

VI. PERFORMANCE ANALYSIS WITH SIMILAR WORKS

Table VI presents a comparison between our suggested models and some current publications. The table demonstrates how classification methods were used in every study to identify OC. On the other hand, we have used twelve machine learning techniques. Additionally, we have utilized SMOTE and multiple imputation methods to address imbalanced and missing data. We have achieved more than 2.3% of accuracy compared with existing approaches. We have evaluated all the five metrics for judging the model performance compared to the existing approaches.

VII. CONCLUSION

We have developed an ensemble-based method for OC detection in this paper. In this case dataset consists of 349 patients with 49 features where 171 are diagnosed with OC and 178 are diagnosed with benign. To improve the model's performance, twelve ML techniques were employed to categorize OC, and key parameters were chosen by fine-tuning the hyperparameters. Out of twelve algorithms, the voting classifier outperformed the others. To address the missing value and unbalanced data, MICE imputation and borderline SVMSMOTE algorithms were also applied. Two feature selection techniques such as Boruta and RFE have also been used. The Boruta method, which produced an average of 93.44% AUC-ROC, 88.57% precision, 93.06% accuracy, 92.54% F-measure, and 96.88% recall, decreased 50% of the total features. Experiments show that our proposed method can recognize OC more correctly than the state-of-the-art methods.

A. Limitations

1) Limited dataset size: The dataset has a constrained number of samples relative to the complexity of ovarian cancer subtypes. This limits the ability to robustly train and validate the model, especially for rare variants.

2) No Prospective clinical validation: The model was tested on retrospective data only. No prospective testing was conducted to assess clinical usability and patient impact.

B. Future Work

1) Multimodal feature integration: Fuse imaging, genomic, and clinical metadata with biomarker data to improve predictive power and facilitate subtype-specific detection.

2) Explainability benchmarking: Compare SHAP with other interpretability techniques like LIME, Integrated Gradients, and Counterfactual Explanations to assess transparency efficacy across user types (clinicians, data scientists).

3) Personalized risk stratification: Expand the fuzzy system to include personalized thresholds for pre- and post-menopausal patients, incorporating age, hormone levels, and family history.

Data Availability

All the data used in this study is publicly available at https://data.mendeley.com/datasets/th7fztbrv9/11.

Acknowledgment

The authors would like to thank Dr. P. Maragathavalli, the research supervisor, for her continuous guidance, support, and encouragement throughout the course of this study. The authors also acknowledge the open data provided by the Mendeley Data platform, which greatly facilitated this research.

References

[1] P. Wu, Q. Jiang, L. Han, and X. Liu, "Systematic analysis and prediction for disease burden of ovarian cancer attributable to hyperglycemia: a comparative study between China and the world from 1990 to 2019," *Front. Med.*, vol. 10, p. 1145487, 2023, doi: 10.3389/fmed.2023.1145487.

[2] Z. Maleki *et al.*, "Survival rate of ovarian cancer in Asian countries: a systematic review and meta-analysis," *BMC Cancer*, vol. 23, p. 558, 2023, doi: 10.1186/s12885-023-11041-8.

[3] P. Dss Lakshmi Kumari and P. Maragathavalli, "Integrating machine learning algorithms for accurate ovarian cancer diagnosis: a state-of-the-art review," in *2023 IEEE ICACIC*, Puducherry, India, 2023, pp. 1–6, doi: 10.1109/ICACIC59454.2023.10434947.

[4] M. Lu *et al.*, "Using machine learning to predict ovarian cancer," *Int. J. Med. Inform.*, vol. 141, p. 104195, 2020, doi: 10.1016/j.ijmedinf.2020.104195.

[5] Md. M. Ahamad *et al.*, "Early-stage detection of ovarian cancer based on clinical data using machine learning approaches," *J. Pers. Med.*, vol. 12, p. 1211, 2022, doi: 10.3390/jpm12081211.

[6] W. Huang *et al.*, "Explainable discovery of disease biomarkers: the case of ovarian cancer to illustrate the best practice in machine learning and Shapley analysis," *J. Biomed. Inform.*, vol. 141, p. 104365, 2023, doi: 10.1016/j.jbi.2023.104365.

[7] J. Martínez-Más *et al.*, "Evaluation of machine learning methods with Fourier transform features for classifying ovarian tumors based on ultrasound images," *PLoS ONE*, vol. 14, p. e0219388, 2019, doi: 10.1371/journal.pone.0219388.

[8] X. Yang, M. Khushi, and K. Shaukat, "Biomarker CA125 feature engineering and class imbalance learning improves ovarian cancer prediction," in *2020 IEEE CSDE*, Gold Coast, Australia, 2020, pp. 1–6, doi: 10.1109/CSDE50874.2020.9411607.

[9] M. Akazawa and K. Hashimoto, "Artificial intelligence in ovarian cancer diagnosis," *Anticancer Res.*, vol. 40, pp. 4795–4800, 2020, doi: 10.21873/anticanres.14482.

[10] P. Charkhchi *et al.*, "CA125 and ovarian cancer: a comprehensive review," *Cancers*, vol. 12, p. 3730, 2020, doi: 10.3390/cancers12123730.

[11] X. Hu, Y. Xie, Y. Yang, and H. Jiang, "Group penalized logistic regressions predict ovarian cancer," *Preprint*, 2021, doi: 10.21203/rs.3.rs-1172360/v1.

[12] R. M. Ghoniem, A. D. Algarni, B. Refky, and A. A. Ewees, "Multi-modal evolutionary deep learning model for ovarian cancer diagnosis," *Symmetry*, vol. 13, p. 643, 2021, doi: 10.3390/sym13040643.

[13] A. Sorayaie Azar *et al.*, "Application of machine learning techniques for predicting survival in ovarian cancer," *BMC Med. Inform. Decis. Mak.*, vol. 22, p. 345, 2022, doi: 10.1186/s12911-022-02087-y.

[14] F. H. Juwono *et al.*, "Ovarian cancer detection using optimized machine learning models with adaptive differential evolution," *Biomed. Signal Process. Control*, vol. 77, p. 103785, 2022, doi: 10.1016/j.bspc.2022.103785.

[15] K. M. Boehm *et al.*, "Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer," *Nat. Cancer*, vol. 3, pp. 723–733, 2022, doi: 10.1038/s43018-022-00388-9.

[16] M. C. Comes *et al.*, "An explainable machine learning ensemble model to predict the risk of ovarian cancer in BRCA-mutated patients," *Front. Oncol.*, vol. 13, p. 1181792, 2023, doi: 10.3389/fonc.2023.1181792.

[17] N. Abuzinadah *et al.*, "Improved prediction of ovarian cancer using ensemble classifier and Shapley explainable AI," *Cancers*, vol. 15, p. 5793, 2023, doi: 10.3390/cancers15245793.

[18] V. Aelgani and D. Vadlakonda, "Explainable artificial intelligence based ensemble machine learning for ovarian cancer stratification using electronic health records," *Int. J. Res. Inf. Technol. Comput. Commun.*, vol. 11, pp. 78–84, 2023, doi: 10.17762/ijritcc.v11i7.7832.

[19] N. Agarwal, Deepakshi, J. Harikiran, Y. B. Lakshmi, A. P. Kumar, E. Muniyandy, and A. Verma, "Predictive modelling for heart disease diagnosis: a comparative study of classifiers," *EAI Endorsed Trans. Pervasive Health Technol.*, vol. 10, no. 1, p. e5, 2024, doi: 10.4108/eetpht.10.5518.