[B.Sc. Engg. Thesis]

Ovarian Cancer Prediction Using Whale Optimization-based XGBoost Algorithm

MD. Zeehad Bin Salim Ameer Rahat Ahmed Khan

Electronics and Communication Engineering Discipline
Science, Engineering and Technology School
Khulna University, Khulna 9208
Bangladesh

December 2025

Ovarian Cancer Prediction Using Whale Optimization-based XGBoost Algorithm

This thesis is submitted to the Electronics and Communication Engineering Discipline in partial compliance of the prerequisites for the degree of Bachelor of Science in Electronics and Communication Engineering, abbreviated as, B. Sc. Engg. (ECE).

By

MD. Zeehad Bin Salim Ameer
Student ID: 210911

Rahat Ahmed Khan
Student ID: 210929

Electronics and Communication Engineering Discipline
Science, Engineering and Technology School
Khulna University, Khulna 9208
Bangladesh
December 2025

RECOMMANDATION

This thesis has been submitted to the Electronics and Communication Engineering Discipline of Khulna University in partial compliance of the prerequisites for the degree of Bachelor of Science in Electronics and Communication Engineering, abbreviated as, B. Sc. Engg. (ECE).

Approved By

(Prof. Dr. Uzzal Biswas)

Electronics and Communication Engineering Discipline (**Supervisor)**

Khulna University, Khulna

(Prof. Dr. Md. Mizanur Rahman)

Electronics and Communication Engineering Discipline **(External Member, Chairman & Head)**

Khulna University, Khulna

(Prof. Dr. Abdullah-Al Nahid)

Electronics and Communication Engineering Discipline (**Board Member)**

Khulna University, Khulna

DECLARATION BY AUTHOR

We hereby certify that we are the only authors of the developed thesis with the title "**Ovarian Cancer Prediction Using Whale Optimization-based XGBoost Algorithm"** under the heartfelt directions of our supervisor **Dr. Uzzal Biswas,** Professor of Electronics and Communication Engineering Discipline, Khulna University, Khulna. The thesis is entirely based on our efforts and it doesn't contain any material that had been published in academic degree or non-degree program in any type of language previously. All the sourses which are used in the thesis have been cited properly. This is the true replica of the thesis which includes final revisions. If there is any short of doubt on our thesis about unethical methods, we are responsible for that.

....

MD. Zeehad Bin Salim Ameer

Student ID: 210911

Electronics and Communication Engineering Discipline

Khulna University, Khulna 9208

....

Rahat Ahmed Khan

Student ID: 210929

Electronics and Communication Engineering Discipline

Khulna University, Khulna 9208

ACKNOWLEDGEMENT

All praise and gratitude to the Almighty for giving us the opportunity to work on the thesis which enable us to make some contributions on it. We are extremely delighted and very much thankful to our honorable supervisor Prof. Dr. Uzzal Biswas for his indispensable counselling and thoughtful feedback throughout the research process. His encouragement motivated us to complete our work with proper shape.

We are also very much thankful to our honorary board member Prof. Dr. Abdullah-Al Nahid who provides us with different insightful observations, analytical suggestions and recommendations on our thesis. His guidance truly intensified the content of our work.

We would like to give a special thanks to our external member Prof. Dr. Mizanur Rahman for his insightful feedback and significant suggestions. Also we would like to extend our gratitude to the faculty and the staff of Electronics and Communication Engineering Discipline at Khulna University.

We want to express our gratitude towards the previous researchers and the scientific community as their works provide the foundation of our thesis. We are very much inspired to make some contribution on this field by their utmost dedications to the knowledge.

Also, we are grateful to our parents, family members and friends for their patience, believe and unfaltering love in us. Their unwavering support has helped us to become motivated towards our journey.

ABSTRACT

Cancer, a bunch of diseases which involve unusual growth of the cell with the ability to attack or span to the other parts of the body. Ovarian cancer, a gynecological malignancy which is the seventh most common cancer in women and eighteenth most frequent cancer overall. This cancer is diagnosed at late stage which causes only five years survival rates at ranges from 93% at early stage and 20% at advanced stage. So, to increase the survival rate after suffering from this disease the early diagnosis of this cancer is very much essential. Several traditional and clinical approaches or tests are used to diagnose the ovarian cancer which are costly and require more time. To slove this problem modern technologies like Machine Learning(ML) are being used nowadays. The purpose of our study is to predict the ovarian cancer and enhance the accuracy of the prediction through hyperparameter optimization. In our study, XGBoost algorithm is used for the prediction of the ovarian cancer. Besides metaheuristic algorithm WOA(Whale Optimization Algorithm) is being used for hyperparameter optimization. The algorithms which were proposed in the study were applied in the dataset that is collected from Mendeley Data Repository. This dataset includes 349 samples with 49 features. After hyperparameter tunning 20 features are found stable and used for increasing the accuracy of the model. Using these selected features the accuracy that the model has shown is 89.69%.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

OC	Ovarian Cancer
WOA	Whale Optimization Algorithm
RF	Random Forest
XGBoost	Extreme Gradient Boosting
AdaBoost	Adaptive Boosting
DT	Decision Tree
LR	Logistic Regression
LGBM	Light Gradient Boosting Machine
SVM	Support Vector Machine
KNN	K-Nearest Neighbour
GBDT	Gradient Boosting Decision Tree
CatBoost	Categorical Boosting
PCA	Principal Component Analysis
RFE	Recursive Feature Elimination
MI	Mutual Information
NCI	National Cancer Institute
MLP	Multilayer Perception
XAI	Explainable Artificial Intelligence
CNN	Convolution Neural Network

ANN	Artificial Neural Network
DCNN	Deep Convolution Neural Network
FNN	Feed Forward Neural Network
MRMR	Minimum Redundancy–Maximum Relevance
ROMA	Risk of Ovarian Malignancy Algorithm
SHAP	Shapley Additive Explanation
CA- 125	Carbohydrate Antigen 125
HE4	Human Epididymis Protein 4
PLT	Platelet Count
CA19-9	Carbohydrate Antigen 19-9
AG	Age
ALP	Alanine Aminotransferase
AFP	Alkaline Phosphatase
CEA	Carcinoembriyonic Antigen
CA72-4	Carbohydrate Antigen 72-4
LYM%	Lymphocyte Ratio
NEU	Neutrophil Ratio
AST	Aspartate Aminotransferase
MNP	Menopause
ALB	Albumin
GGT	Gama Glutamyl Transferase
TBIL	Total Bilirubin
HGB	Hemoglobin

IBIL	Indirect Bilirubin
UA	Uric Acid
LYM#	Lymphocyte Count
PDW	Platelet Distribution Width
FS	Feature selection
AUC	Area Under Curve
ROC	Receiver Operating Characterstic
TP	True Positive
TN	True Negative
FP	False positive
FN	False negative

Chapter 1: Introduction

1.1 Background

Ovarian cancer, the most lethal malignancy related to the female reproductive system, is the most usual cancers in women. According to the most recent published data in 2021, there are roughly 314,000 new ovarian 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]. This cancer is typically diagnosed at late stage which causes the high mortality rate due to this cancer [2]. R. L. Seigel et al. noted in their study that, the disease is diagnosed at very late stage for 70% patients thus the survival rate drops to 30-50% [3]. A study supervised by M. K. Hong et al. about the early diagnosis of ovarian cancer remains a significant challenge owing to it's asymptotic nature. Besides some biomarkers including CA-125 and HE4 have shown variety in their sensitivity and specificity [4]. But the faster identification of this disease is a must to enhance the survival rate. With the help of conventional medical testing system it is hard to determine this disease at the early stage. However many advanced technologies like machine learning have been introduced in ovarian cancer research field which increase the survival rate after suffering from this disease [5].

T. Oznacar and T. Guiler [6] trained the machine learning model with catboost classifier along with boruta feature selection method and achieved the accuracy of 89.52% in predicting the ovarian cancer. Several machine learning methods like random forest (RF), xgboost, decision tree(DT), support vector machine (SVM), k-nearest neighbour (KNN) can be used in prediction of ovarian cancer. But without optimization of the model's hyperparameter the accuracy of the prediction becomes lower. That's why hyperparameter optimization is an important part to increase the model accuracy. Several optimization algorithms are used in machine learning. Among them the nature inspired meta heuristic algorithms are very much promising. Besides these algorithms are easy to implement as well. Among all the nature inspired optimization algorithms whale optimization algorithm (WOA) is the most popular [7]. This algorithm works on the basis of the hunting machanism of the humpback whale where the whale generates a spiral to simulate the bubble net attacking. The efficiency of this model developed by solving twenty nine mathematical and six structural optimization problems [8].

WOA is a higher potential optimization algorithm. Though the potential of this algorithm is very high it has some limitations as well. From the simulation outcome it is seen that the algorithm amalgamates a center bias operator which limits it's performance [9]. Some necessary modifications will be taken in this study to solve this problem and enhance the overall performance of the model.

1.2 Motivation

Ovary is the most important part of female reproductive system. The cancer that occurred to this region of the human body is called ovarian cancer. The most common gynecological malignancies found in women which causes early death is ovarian cancer. There are several causes due to early death after suffering from ovarian cancer. Among them the most remarkable point is the advanced stage diagnosis of the disease. Wilson et al.[10] reported in their study that the principal and most important treatment option for the ovarian cancer is surgery. The types and characterstics of the ovarian tumor can only be confirmed after collecting the histopathological samples through the surgery. This time consuming and painful process traumatized the patient. After diagnosis of the disease some further treatments are also required such as chemotherapy where high power anti-cancer drugs are used to destroy the fast growing cancer cell and resurgery to confirm the tumor condition after chemotherapy. All these process leads to high life risk of the patient. Besides it creates financial unstability and mental malfunction to the patient as well as to their families. Detecting the ovarian cancer with the help of conventional medical system is very much time consuming and disturbing. In this regard to solve this problem several studies have done where with the aid of several machine learning algorithms the ovarian cancer will be detected at the early stage. In recent time, machine learning has found widespread applications [11]. Several machine learning techniques are being used in the prediction of ovarian cancer. Among them XGBoost have shown an incredible performance. A study presented by Ozhan et al.[12] to detect the ovarian cancer using machine learning algorithm where they built a model with the XGBoost classifier. The performance of the model is 89.5% which is a good rate to detect the ovarian cancer. The increament of the efiiciency of a model is dependent on hyperparameter tuning. Every model has some different hyperparameters. To tune the hyperparameter with the aim of increasing the efficiency of the model with respective dataset metaheuristic optimization algorithm is required. Whale optimization algorithm (WOA) is a natural metaheuristic algorithm which can be used for the optimization. Though it has some limitations some modifications will be made to mitigate this problem. Thus the increament of the model's accuracy will make the model more reliable in the detection of the ovarian cancer. So, we have been inspired to generate a model which will detect the ovarian cancer and enhance the accuracy as well as make the model more reliable. Thus different clinicians can use this model to detect the ovarian cancer early and diminish the above mentioned difficulties.

1.3 Problem Statement

The advanced stage detection in most women of the ovarian cancer is caused by the poor prognosis [13]. Several challenges are being faced due to the detection of this disease. The major challenge is the late detection of this disease. The survival rate after suffering from this disease is not so high. This disease is diagnosed due to the metastatic disease in the pelvis in almost 90% patient and the survival rate for five years of these patients are less than 30%. Secondly, the identification test with enough sensitivity for ovarian cancer is a challenge. Maximum markers are generated using the sample taken from the patient. But the problem arise when it is needed to be identified not only the ovarian cancer but also early disease before it causes symptoms [14]. Extracting the most significant features from the enormous patient data is also a big challenge. This may cause inappropriate prediction. Medical dataset contains huge number of variables and all these variables are not so much relevant. Using feature selection method will select the important features relevant to the prediction test. The productivity of different machine learning models depend on the effective hyperparameter tuning. Among all the model XGBoost have shown great commitment in predicting ovarian cancer using effective hyperparameter tuning with whale optimization algorithm (WOA). The main reason of selecting this optimization algorithm is that it can find the optimal solution with fast speed. Though this algorithm has some limitations, some necessary measures will be taken to mitigate this problem and enhance the overall accuracy of the model.

1.4 Objectives of the Thesis

This study persues to predict the ovarian cancer appropriately. To achieve this goal the given objectives has been set for the research.

- Investigating the ovarian cancer dataset and run this dataset into 3 different models named Random Forest, XGBoost and CatBoost to find the best result of predicting the ovarian cancer. Among them XGBoost gave the best result.

- Extracting the important feature using feature selection method to predict the ovarian canacer.
- Optimizing the hyperparameter of XGBoost using whale optimization algorithm(WOA).
- Performance exploration of the optimized XGBoost model for the ovarian cancer prediction.
- Making comparison between the result of our study with the related works.

1.5 Contribution of the Thesis

At the termination of our study, we have effectively developed a model with an intensified ability to predict the ovarian cancer. With the help of our study, we have tried to make small contributions to the patients who have suffered from this disease. We have tried to enhance the ability of our model as much as possible so that it can give more precise result. Thus doctors can guide the risky patient and the patient with low risk according to the result of the developed model. Using this model will diminish the difficulties that patients are facing in conventional medical system. It will reduce the unnecessary medical check up as well. Our study have forwarded the uses of advanced technologies like machine learning in the field of healthcare which will create great impact in the medical sector in near future.

1.6 Organization of the Thesis

This thesis has been partitioned into several chapters. The zest of the all the chapters in the study are described below at a glance.

Chapter 1

In the introduction chapter, there is an overview including the background, motivation, problem statement, objectives and the contribution of the thesis has been provided.

Chapter 2

In the literature review chapter, there is a systematic review of our related study has been provided. The literature review of our study is divided into three major sectors. The related papers with those major sectors has been reviewed systematically in this chapter. It addition the chapter also include different findings which measure the result of different models of the related work.

Chapter 3

A compact overview of our study’s methodology, dataset selection, model selection, feature selection method, hyperparameter tuning and the performance analysis metrices has been provived in this chapter.

Chapter 4

The chapter result and discussion provides the detail of our model performance’s using whale optimization algorithm (WOA). It also provides the summary of feature selection, hyperparameter tuning of the model. Finally, it demonstrates the performance comparison with the related works.

Chapter 5

The summary and the major key findings of our study has been demonstrated in the chapter of conclusion and future work. Besides it also provides a comprehensive knowledge over the future research of our study and the related areas where our proposed model can be improved.

Chapter 2: Literature Review

2.1 Introduction

Relevant literature is very much important in the field of all research works. While studying an article the author starts to describe previous researches and try to map and assess the research to motivate the aim of the study and justify the hypothesis which is considered as the literature review. Different approaches are used in literature review. They are the systematic approach, semi-systematic approach, bibliometric review and others [15]. Among these approaches for our work, we have implemented the systematic approach. The objective of this process is to determine all empirical evidence that fits the pre specified inclusion criteria to answer a particular research hypothesis. Due to the above mentioned reasons the systematic approach has been used in this study.

2.2 Systematic Literature Review

Systematic literature review is a way of consolidating scientific evidence to answer a specific research query in a way that is clear and reconstructable, while seeking to include all published evidence on the topic and appraising the quality of the evidence. This review process has become an important methodology in different sectors including public policy research and health sciences [16]. Systematic reviews combine results across different studies and determine an overall effect [3]. To perform systematic analysis we have organized our work into three important parts reffered as feature selection, hyperparameter tuning and classification.

2.2.1 Feature Selection

Feature selection is an effective way to reduce dimensionality by eliminating irrelevant data. Most traditional feature selection approaches score and rank each feature individually and then perform feature selection by eliminating lower ranked features or by retaining higher ranked features [18]. So, selection of the feature is an important part in the research region. Thus we have studied several papers where different feature selection methods are applied to select the relevant features to detect the ovarian cancer. Among them some important papers review and their findings using feature selection method are given below.

D. Lakshmikumari and P. Maragathavalli [19] presented comparative analysis of machine learning frameworks for robust ovarian cancer detection using feature selection and data balancing based on voting classifier with Boruta. The main purpose of the study was to detect the ovarian cancer using certain clinical examination. Several ensemble techniques including random forest, gradiant boosting decision tree (GBDT), adaptive boosting, bagging, xg-boost were used here. For the selection of the feature, boruta feature selection method was used. Using feature selection method 21 features were selected among 24 and these were AFP, age, ALB, ALP, AST, CA125, CA19-9, CEA, GEO, HE4, IBIL, LYM#, LYM%, MCH, Menopause, MPV, Na, PCT, PLT, TBIL and TP. By using voting classifier with boruta gave the highest average of accuracy 93.06%, precision 88.57%, recall 96.88%, f1-score 92.54% and AUC-ROC base result 93.44%. 2025

T. Oznacar and T. Guiler [6] demonstrated in the study about the early diagnosis of ovarian cancer in patients using machine learning approaches with boruta and advanced feature selection where they used eight machine learning techniques such as Random Forest, XGBoost, CatBoost, Decision Tree, K-Nearest Neighbors, Naïve Bayes, Gradient Boosting and Suppport Vector Machine. Besides four feature selection method like Boruta, PCA, RFE and MI were also used for selecting appropriate feature for finding maximum efficiency of the model. Among all other calssifiers catboost's with boruta feature selection method showed the maximum result. With the help of a feature selection method called boruta, a set of 20 features were found and these were AFP, age, ALB, ALP, AST, CA125, CA19-9, CA72-4, CEA, GLO, HE4, IBIL, LYM#, LYM%, MCH, NEU, Na, PCT, PLT, TBIL and TP. The maximum result were accuracy 89.52%, f1-Score 89.46%, precision 90.73%, recall 89.52% and AUC 95.03%. In the research three biomarkesrs named as HE4, CA125 and CA72-4 exhibited great impact on the model's predictive outcome.

A study presented by J. Cai et al. [20] about integrated algorithm where they used feature selection, data augmentation and xgboost for ovarian cancer prediction and the purpose of their study was to develop a model to predict the ovarian cancer at an early stage utilizing genomic data. To increase the model's predictive accuracy the original genetic dataset was simplified through feature selection method. To classify the augmented data XGBoost classifier was applied in the research. The accuracy that shown by the model through xgboost classifier was 99.01% and twelve genes were found by the model which were highly relevant to ovarian cancer and they were FTCD, KRTAP5-8, GSK3A, RAB5A, SNAPC5, CHTOP, GPR137, GNAQ, UBAP1, IGHMBP2, TSPAN15, SMPD3. By using embedded feature selection method (random forest) 8 features or genes were found and these were FTCD, KRTAP5-8, GSK3A, RAB5A, SNAPC5, CHTOP, GPR137, GNAQ.

B. Wickramasinghe and G. Regisford [21] presented a comparative study to predict ovarian cancer using three types of machine learning algorithm such as, Logistic Regression, Decision Tree and KNN. The dataset that had been used in the paper had taken from the Kaggle online data repository, representing 335 instances and 49 features. For finding the appropriate features, a feature selection method called chi-square test was applied in the model. After applying feature selection method 21 features out of 49 were found and these were AFP, AG, ALB, ALP, AST, CA125, CA19-9, CA72-4, CEA, GGT, HE4, IBIL, LYM%, LYM#, UA, HGB, MNP, NEU, PLT, TBIL, and PDW. Here in the paper basically two performance indiactors were used and these were accuracy and recall. In the paper Logistic Regression, Decision Tree and KNN were used in three different splitting ratio(20%-80%, 30%-70%, 40%-60%) of training and testing data. Among all the algorithms logistic regression gave the maximum accuracy with 87% and recall of 99%.

A study conducted by T. Gui et al. about the early prediction and risk stratisfaction of ovarian cancer using several machine learning methods including SVM, Bayes, LR, DT, Light GBM and XGBoost. The dataset consisted of 9799 patients. Using feature selection method 27971 dimensional features were extracted from 7455 patients. Then these features were filtered and 663 features were obtained. Five fold cross validation were applied for each model and 20 features were ranked bye shapely additive interpretation(SHAP) method. Age, CA125 and risk of ovarian malignancy were the top three feature out of 20 features. Among all the models light GBM gave the maximum result of accuracy 88.29%, recall 87.66%, precision 87.65% and f1-score 87.86%.

L. Akter and N. Akhter [22] demonstrated prediction of ovarian cancer based on TVUS using machine learning approach. In their study, they used three machine learning techiniques such as Random Forest, XGBoost and KNN. The dataset that was used in the study was collected from National Cancer Institute (NCI), United States. KNN Imputer was used to handling the missing values from the dataset. Feature scaling method called min max normalization was used to bringing most of the features to the same scale. 18 features were scaled and they were numcystl, numcystr, ovary_diaml, ovary_diamr, ovary_voll, ovary_volr, ovycyst_diaml, ovycyst_diamr, ovyyst_morphl, ovyyst_morphr, ovyyst_outlinel, ovyyst_outliner ovyyst_solidl, ovyyst_solidr, ovyyst_suml, ovcyst_sumr, ovcyst_voll, ovcyst_volr. This model gave accuracy 99.50%, recall 99.50%, f1-score 99.50% and precision 99.49

2.2.2 Hyperparameter Tuning

Hyperparameter Tuning is the process of determining the best possible hyperparameters. This process is used to develop several tools to traverse the space of possible hyperparameter configurations systematically and in organized way [23]. In machine learning there are several types of model containing different hyperparameters. These hyperparameters may get changed for different types of data. So it is important to make the model stable by tuning the hyperparameters properly. For our study hyperparameter tuning is an important stage. So, we have studied several papers where hyperparameter tuning is used to enhance the model stability. Among them the important and relevant papers review has given below.

R. Shetty et al. [24] presented the optimization of the machine learning based ovarian cancer prediction through normalization strategies where they basically used clinical and biomarker-based data. Two datasets were used in the study. The data was scaled through min-max scaling and z-score normalization. Several machine learning algorithms were used such as KNN, LR, SVM, MLP, Stacking, Bagging. In this study, random forest classifier using gini impurity was used to assess the feature selection. After hyperparameter tunning several hyperparameters were found for different classifiers. Among them the following hyperparameters like KNN, SVM C and Meta-Model C were found in stacking classifier. In this study, stacking gave the maximum accuracy of 100%. For dataset 1, the precision, recall and f1-score were also 100% but for dataset 2, the precision, recall and f1-score differs with the different class of the data.

H. Dhingra and R. Shetty [25] demonstrated a study where they analyzed comparatively between machine learning and deep learning models for the early prediction of ovarian cancer using clinical and biomarker data. Several machine learning classifiers including SVM, KNN, LR, RF were used. For deep learning networks like ANN, RNN, FNN and CNN were evaluated. Several ensemble techniques including stacking, bagging adaboost and xgboost were also used. For feature selection several methods like feature importance, recursive feature elimination (RFE) and autoencoder based techniques were used. Hyperparameter tunning and optimization of the model were done to enhance the model performance. Adam optimizer was used in the optimization process of the model. Among all the classifiers and networks FNN combined with autoencoder based feature selection techniques achieved the highest accuracy of 85.71% and the tuned hyperparameters for the FNN model were Optimizer $= \text{Adam}, \text{Epochs} = \text{fixed standardized value}, \text{Batch}$ size $= \text{fixed standardized value}$.

A study conducted by A. S. Azar et al. [26] about the applications of machine learning techniques to predict the ovarian cancer. In their study, they used several machine learning techniques to predict the ovarian cancer such as SVM, LR, KNN, RF, DT, AdaBoost, XGBoost. For feature selection, min-max normalization method was used. To create the association between the features pearson correaltion coefficient was used. In the case of hyperparameter tunning, grid search was used. After hyperparameter tunning the following hyperparameter such as criterion: gini, max_depth: none and n_estimators: 100 was found for random forest classifier. Among all the classifiers random forest showed the maximum accuracy 88.72% and AUC 82.38 %.

2.2.3 Classification

Classification is a supervised learning technique that includes categorizing data into distinct classes. It is used to predict the result of a given problem based on input features. In machine learning classification is very much important to categorize the data into distinct classes. To get the accurate result from the model classification is necessary. The following papers are related with the classification of different machine learning models.

Y. Sun and B. Wen [27] presented machine-learning diagnostic model for ovarian tumors where they used four different types of machine learning algorithm such as Random Forest, K-Nearest Neighbors, Suppport Vector Machine and Logistic Regression. 713 patients with ovarian tumors at Sun Yat Sen Memorial Hospital were randomized into training and testing cohorts. To reduce unnecessary features, a feature selection method [28] called sequential backward selection (SBS) was used. Among all the classifier random forest showed the highest accuracy of 99.82% with micro average ROC of 0.86 for the malignant ovarian cancer from benign or borderline ovarian tumors and for pathological tissue logistic regression shows accuracy of 78% and micro-average ROC curve 0.95.

S. L. J. M and S. P demonstrated [29] in a study about the creative approach towards the faster prediction of ovarian cancer using machine learning-enabled XAI techniques where they used machine learning techniques along with explainable artificial intelligence(XAI) to predict the ovarian cancer at an early stage. In this research several machine learning techiniques were used such as k-nearest neighbors, suppport vector machine, decision tree and ensemble learning techniques sucha as max vooting, boosting, bagging and stacking. The dataset that was used in the research was taken from Mendeley Dataset Repository. support vector machine showed accuracy of 85% with the base model whereas 89% accuracy was achieved after stacking.

A study conducted by S. D. Patil , P. J. Deore and V. B. Patil [30] about an intelligent computer aided diagnosis system for classification of ovarian masses using machine learning approach where the objective of the research was to detect the ovarian mass utilizing novel annotated ovarian masses. For extracting the feature, several feature extraction techniques were used such as gray level co-occurrence matrix(GLCM), gabor, edge and tamura. Binary segmentation with RF classifier gave the maximum accuracy of 86% where as KNN gave accuracy below 84%.

S. M. Ayyoubzadeh et al.[31] demonstrated the ovarian cancer prediction using different artificial tools. The purpose of their study was to predict the ovarian cancer according to the characterstics and conditions of each person. The dataset used in their study consisted of 349 patients with 49 features. 171 patients out of 349 were related to benign tumors and rest of the others were malignant tumors . Four different models with ten fold cross validations were used in the study. The model used in the study were random forest (RF), decision tree (DT), support vector machine (SVM) and artificial neural network (ANN). Among all the models random forest gave the maximum result with accuracy 86.75%, AUC 92.5%, f1-score 88.01%, sensitivity 91.60% and specificity 81.36%.

A study presented by M. Lu et al.[32] about to predict the ovarian cancer using three models decision tree (DT), logistic regression (LR) and risk of ovarian malignancy algorithm (ROMA). For the feature selection from the dataset MRMR (Minimum Redundancy–Maximum Relevance) algorithm was used. MRMR is a filter type feature selection method proposed by C. Ding and H. Peng [33]. The risk of ovarian malignancy algorithm was proposed by Moore et al [34]. Among all the model logistic regression gave the maximum accuracy of 97.4%, precision 92.3%, recall 96% and specificity 97.8%.

A. Arfiani and Z. Rustam [35] presented a study over the classification of ovarian cancer data using bagging and random forest. In this paper, the dataset that was used had taken from the UCI Machine Learning Repository. Number of samples that were used in this classification was 266 with 15154 feature. Here 10% data was used as test data and 90 % was used for training data. For the classification of ovarian cancer different types of methods were used, support vector machine (SVM) was one of them. Other approaches were also used like clinical and integrative method [36]. But in this paper bagging and random forest were used for classification. In bagging method, 100% accuracy was seen with run time 0.164178 . In random forest method, the accuracy was less and it was 98.2% with run time 0.127821.

Table 2.1: Summary Of Literature Review

Literature Title	Dataset	Authors Name	Accuracy	Precision	Recall	Specificity	F1-Score
Comparative Analysis of Machine Learning Frameworks for Robust Ovarian Cancer Detection Using feature Selection and Data Balancing	Sample : 349 Feature; 49 Benign:178 Malignant:171	D. Lakshmikumari and P.Maragathavalli	93.06%	88.57%	96.88%		92.54%
Prediction of Early Diagnosis in Ovarian Cancer Patients Using Machine Learning Approaches With Boruta and Advanced Feature Selection	Sample : 349 Feature: 49 Benign:178 Malignant:171	T. Oznacar and T. Guiler	89.52%	90.73%	89.52%		90.73%

Literature Title	Dataset	Authors Name	Accuracy	Precision	Recall	Specificity	F1-Score
An Integrated Algorithm With Feature Selection, Data Augmentation and XGBoost for Ovarian Cancer	Sample:502 Feature: 11476	J. Cai et.al	99.01%				
A Comparative Study to Predict Ovarian Cancer	Sample:335 Feature: 49 positive(1): 171 negative(0): 164	B. Wickramasinghe and G. Regisford	87% (Logistic Regression)		99%		
An Integrated Algorithm with Feature Selection, Data Augmentation and XGBoost for Ovarian Cancer	Sample : 349 Feature; 49	T. Gui et al.	99.01%				
Ovarian Cancer Prediction from Ovarian Cysts Based on TVUS Using Machine Learning Approach	Feature: 18	L. Akter and N. Akhter	99.5%	99.5%	99.5%		99.49%

Literature Title	Dataset	Authors Name	Accuracy	Precision	Recall	Specificity	F1-Score
Optimizing Machine Learning Based Ovarian Cancer Prediction Through Normalization Stratagies	Sample:30989 Feature: 134	R. Shetty et al.	100% (dataset -1)	100% (dataset -1)	100% (dataset -1)		100% (dataset -1)
Comparative Study of Machine Learning and Deep Learning Models for Early Prediction of Ovarian Cancer	Sample:349 Feature :49 Benign:178 Malignant:171	H. Dhingra and R. Shetty	85.71%				
Applications of Machine Learning Techniques for Predicting Survival in Ovarian Cancer	Sample:42827 Feature :17	A. S. Azar et al.	88.7%				71.7%
Machine-learning Diagnostic Model for Ovarian Tumors	Sample:713 Benign:304 Borderline:94 Malignant:311	Y. Sun and B. Wen	99.82% with AUC 0.86				

Literature Title	Dataset	Authors Name	Accuracy	Precision	Recall	Specificity	F1-Score
Innovative Approach Toward Early Prediction of Ovarian Cancer: Machine Learning-Enabled XAI Techniques	Sample : 349 Feature :51 Benign:171 Malignant:176	S. L. J. M and S. P	89%				
An Intelligent Computer Aided Diagnosis System for Classification of Ovarian Masses Using Machine Learning Approach	Sample : 187 Benign:112 Malignant:75	S. D. Patil , P. J. Deore and V. B. Patil	86%				
Prediction of Ovarian Cancer Using Artificial Intelligence Tools	Sample:349 Feature :49 Benign:178 Malignant:171	S.M. Ayyoubzadeh et al.	86.75%		91.6%	81.36%	88.01%
Using Machine Learning To Predict Ovarian Cancer	Sample:349 Feature :49 Benign:178 Malignant:171	M. Lu et al.	97.4% (Logistic Regression)	92.3%	96%	97.8%	

Literature Title	Dataset	Authors Name	Accuracy	Precision	Recall	Specificity	F1-Score
Ovarian Cancer Data Classification Using Bagging and Random Forest	Sample:266 Feature:15154	A.Arfiani and Z.Rustam	100%(Bagging)**,** 98.2%(RF)				

2.3 Discussion

Ovarian cancer, the most common malignancy related to the female reproductive system is the most lethal cancers in women. According to the most recent published data in 2021, there are roughly 314,000 new ovarian 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). To get remission from this disease the faster detection of the ovarian cancer is a must. This systematic analysis provides a great understanding over different techniques used by different authors to develop a model to predict the ovarian cancer. To disentangle our analysis and share a brief information, we have divided our analysis into three parts. And these parts are feature selection, hyperparameter tuning and classification. The first among three parts are feature selection. In this part of the study, D. Lakshmikumari and P. Maragathavalli [19] used boruta feature selection method to extract 21 features out of 49. The extracted features were AFP, age, ALB, ALP, AST, CA125, CA19-9, CEA, GEO, HE4, IBIL, LYM#, LYM%, MCH, Menopause, MPV, Na, PCT, PLT, TBIL and TP. In another study, T. Oznacar and T. Guiler [6] used the same dataset and same feature selection technique as the previous authors but they extracted 20 features. And the features were AFP, age, ALB, ALP, AST, CA125, CA19-9, CA72-4, CEA, GLO, HE4, IBIL, LYM#, LYM%, MCH, NEU, Na, PCT, PLT, TBIL and TP. These three features CA72-4, GLO, NEU were not found in the study of the previous authors. In another study B. Wickramasinghe and G. Regisford [21] also extracted 21 features using the same dataset.

But there were some differences in the extracted features. The extracted features were AFP, AG, ALB, ALP, AST, CA125, CA19-9, CA72-4, CEA, GGT, HE4, IBIL, LYM%, LYM#, UA, HGB, MNP, NEU, PLT, TBIL, and PDW. HGB, GGT, PDW were not found in the previous authors study. J. Cai et al. [20] an another author used feature selection and data augmentation method to extract the important features from the dataset. But in their study they used genomic data and using feature selection method to extract 8 features out of 12 and these features were FTCD, KRTAP5-8, GSK3A, RAB5A, SNAPC5, CHTOP, GPR137, GNAQ. In the study of the early prediction and risk stratisfaction of ovarian cancer using several machine learning methods T. Gui et al. extracted 20 features using SHAP. Among these features. Age, CA125 and risk of ovarian malignancy (ROMA) were the top three. In the study of predicting the ovarian cancer depended on TVUS using machine learning approach, L. Akter and N. Akhter [22] scaled the features using min max normalization method and the features were numcystl, numcystr, ovary_diaml, ovary_diamr, ovary_voll, ovary_volr, ovycyst_diaml, ovycyst_diamr, ovyyst_morphl, ovyyst_morphr, ovyyst_outlinel, ovyyst_outliner ovyyst_solidl, ovyyst_solidr, ovyyst_suml, ovcyst_sumr, ovcyst_voll, ovcyst_volr.

The second most important is the hyperparameter tuning. In the study of the optimization of the machine learning based ovarian cancer prediction through normalization strategies R. Shetty et al. [24] used min-max scaling and z-score normalization. After the process of hyperparameter tuning they found the following hyperparameters KNN, SVM C and Meta-Model C. In another study H. Dhingra et al. used several machine learning classifiers including SVM, KNN, LR, RF and deep learning networks like ANN, RNN, FNN and CNN to demonstrate the comparative analysis of machine learning and deep learning models for early ovarian cancer prediction using clinical and biomarker data. In their study they got max accuracy with FNN and the hyperparameters were Optimizer= Adam,Epochs = fixed standardized value, Batch size = fixed standardized value.

In the study of the applications of machine learning techniques to predict the ovarian cancer conducted by A. S. Azar et al. [26] demonstrated several machine learning techniques such as SVM , LR, KNN, RF, DT, AdaBoost, XGBoost to predict the ovarian cancer. In the case of hyperparameter tunning, grid search was used. After hyperparameter tunning the following hyperparameter such as criterion: gini, max_depth: none and n_estimators: 100 was found for random forest classifier.

The third and final part of this review process is classification. In the study of machine-learning diagnostic model for ovarian tumors Y. Sun and B. Wen [27] used four machine learning algorithm Random Forest, K-Nearest Neighbors, Suppport Vector Machine and Logistic Regression. Among all the classifier random forest showed the highest accuracy of 99.82% with micro average ROC of 0.86 for the malignant ovarian cancer from benign or borderline ovarian tumors and for pathological tissue logistic regression shows accuracy of 78% and micro-average ROC curve 0.95. In another study S. L. J. M and S. P demonstrated [29] machine learning-enabled XAI techniques such as K-Nearest Neighbors, Suppport Vector Machine, Decision Tree and ensemble learning techniques such as Max Vooting, Boosting, Bagging and Stacking to predict the ovarian cancer at early stage. Among all the classifiers Support Vector Machine showed accuracy 85% with the base model whereas 89% accuracy was achieved after stacking. S. D. Patil , P. J. Deore and V. B. Patil

[30] presented an intelligent computer aided diagnosis system for classification of ovarian masses using machine learning approach where Binary segmentation with RF classifier showed the maximum accuracy of 86% where as KNN gave accuracy below 84%. In another study, named the prediction of ovarian cancer using different artificial tools demonstrated by S. M.

S. M. Ayyoubzadeh et al. [31] used random forest (RF), decision tree (DT), support vector machine (SVM) and artificial neural network (ANN). Among all the models random forest showed the maximum result with accuracy 86.75%, AUC 92.5%, f1-score 88.01%, sensitivity 91.60% and specificity 81.36%. Another authors named M. Lu et al. [32] presented using machine learning to predict ovarian cancer using three models DT(Decision Tree), LR(logistic Regression) and ROMA(Risk of Ovarian Malignancy Algorithm). Among all the model Logistic regression gives the maximum accuracy of 97.4%, precision 92.3%, recall 96% and specificity 97.8%. In the study, of ovarian cancer data classification using bagging and random forest demonstrated by A. Arfiani and Z. Rustam [35] used different types of methods. Among them in bagging method 100% accuracy was seen with run time 0.164178 . In random forest, method the accuracy was less and it was 98.2% with run time 0.127821.

2.4 Conclusion

The systematic analysis emphasizes significant upgradation in ovarian cancer prediction using several machine learning approaches. In this study we have demonstrated three important parts related to the prediction of the ovarian cancer and these are feature selection, hyperparameter tuning and classification. The application of machine learning in the field of the detection and the prediction of the ovarian cancer enhance the chances to track and predict the disease at an early stage over the conventional detection techniques. Furthermore this study will give the authors and researchers with a comprehensive outlook over different sectors of research over the field of ovarian cancer prediction and detection.

Chapter 3: Materials and Methods

3.1 Introduction

This chapter illustrates the materials and methods of our study. The primary objective of our work is to build a model using robust classifier along with a metaheuristic algorithm that can accurately and efficiently predict ovarian cancer. A systematic workflow was followed to complete our work.

The flow was:

- Dataset acquisition
- Dataset Preprocessing
- Evaluation of different classifiers
- Feature Selection using WOA
- Hyperparameter tuning
- Model training
- Performance analysis

Multiple machine learning classifiers were used and evaluated and XGBoost was selected for ultimate research. Important features were selected using Whale Optimization Algorithm and hyperparameter tuning was incorporated to optimize the model further for improved metrics. Each of these steps are described in the subsequent sections to demonstrate the methodology applied in our study.

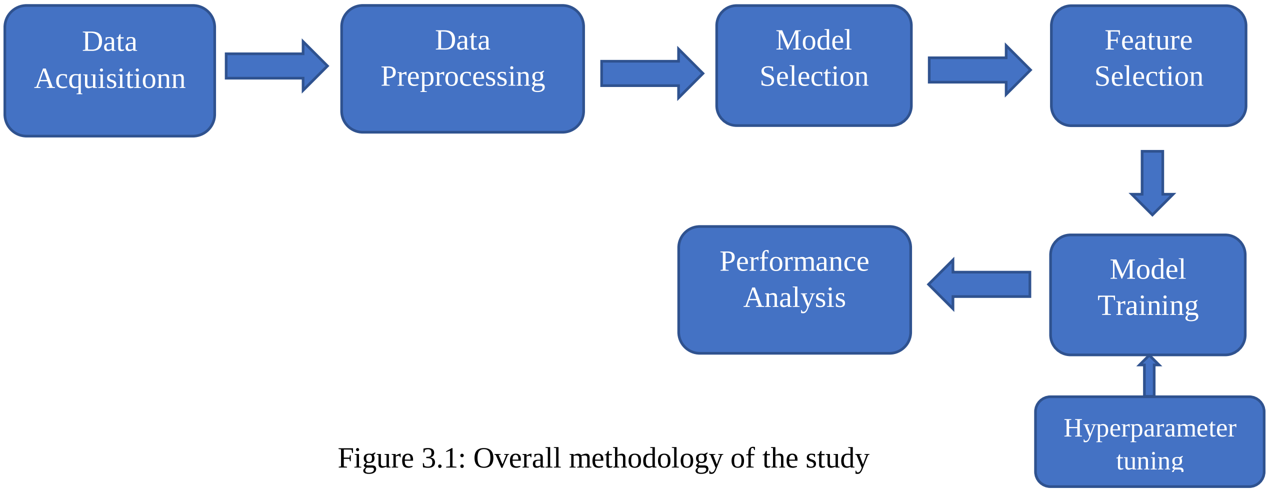

Figure 3.1: Overall methodology of the study

3.2 Data Acquisition

We collected the dataset from a publicly available ovarian cancer repository from Mendeley Data platform. The dataset has clinical information of 349 patients which was collected from the 'Third Affiliated Hospital of Soochow University' [32]. It was collected from 'July 2011 to July 2018', and the number of ovarian cancer patients were 171 and the number of patients who have benign ovarian tumor were 178. This made the dataset a balanced. It has a feature set of 49 features. Pathology diagnosis was conducted to collect the features [37]. The dataset consists of clinically relevant biomarkers which are crucial for predicting ovarian cancer. The dataset has ethical approval which ensures that the information of the patients were collected by following proper institutional and ethical guidelines[32]. Below The dataset is divided into three subgroups in table 3.1

Table 3.1: Different subgroup's attribute list of the dataset

General Chemistry	**Blood Routine Test**	**Tumor Marker**
Albumin	Neutrophil ratio	Carbohydrate antigen 72-4
Indirect bilirubin	Thrombocytocrit	Alpha-fetoprotein
Uric acid	Hematocrit	Carbohydrate antigen 19-9
Nutrium	Mean corpuscular	Menopause
Total protein	hemoglobin	Carbohydrate antigen 125
Alanine aminotransderase	Lymphocyte	Carcinoembryonic antigen
Total bilirubin	Platelet distribution width	Age
Blood urea nitrogen	Mean corpuscular volume	Human epididymic protein 4
Magnesium	Platelet count	
Glucose	Hemoglobin	
Creatinine	Eosinophil ratio	
Phosphorus	Mean platelet volume	
Globulin	Red blood cell count	
Gama glutamyl tranferasey	Mononuclear cell count	

Alkaline phosphates	Red blood cell distribution width	
Kalium	Basophil cell ratios	
Direct bilirubin		
Carbon dioxide-combining power		
Chlorine		
Aspartate aminotransferase		
Anion gap		

3.3 Data Preprocessing

Missing values, noise and outliers made raw data improper and unsuitable for analysis which leads to erroneous results. Data preprocessing handles this issue. It enhances the efficiency and accuracy of the data by using different effective preprocessing techniques [38]. Our collected dataset was investigated and found some missing values in several crucial biomarker variables. Hence a simple mean-imputation technique was used to handle the missing entries where we replaced the missing values with the meaning of its corresponding feature. It prevented the biasness because of the lost samples and preserves the overall data distribution. As our collected dataset has numerical values only, no categorical encoding was required. We used tree-based classifiers for our research so no scaling or standardization were applied as they are invariant to feature scaling.

3.4 Model Evaluation:

After preprocessing the data, we aimed to select a best performed classifier with the dataset. We used three machine learning classifiers which are Random Forest, CatBoost and XGBoost to measure the predicting capability of ovarian cancer.

Random Forest: The classifier is made of many independent classifiers or decision trees. It decides the class label of an input sample using the voting results of each decision tree. [39] . It is based on two main ideas:

1. Bootstrap Aggregation (Bagging): Bootstrap sample is pick from the dataset with replacement from the dataset and each tree is trained on that sample.
2. Random Feature Subsampling: Only a arbitrary subset of features is picked from the dataset at each tree split which improves the adaptability of the model and decorelates them.

XGBoost: It is a complex gradient boosting framework which increases the computational speed and accuracy of prediction. XGBoost differs from Random Forest in the context that it builds trees sequentially whereas RF trains trees independently. Each tree fixes the error outputted by the previous tree in XGBoost. XGBoost is highly effective for biomedical classifications as it shows consistent achievement of state-of-the-art performance where the data is structured.

CatBoost: CatBoost is a type of boosting algorithm which was developed to minimize some of the problems of conventional gradient boosting such as overfitting and sensitivity. The working of catboost is based on the gradient boosting but handling categorical data efficiently and smartly makes it special.

3.5 Feature Selection

Feature selection is a technique used for finding the best subset of features form a feature set. Through which generalization error can be minimized. The machine learning and data mining community is using feature selection comprehensively [40]. In our study one of our goals was to achieve effective accuracy while reducing the feature. We used Whale Optimization Algorithm for this task. This metaheuristic algorithm overcomes the optimization problems by following the biological or physical phenomena [8]. The algorithm obtains the optimal solution by efficiently driving in a broad search space while maintaining the balance between the exploration and exploitation phases. This optimization algorithm becomes more valuable because of its less computational overhead to deliver reliable results [41].

3.5.1 Whale Optimization Algorithm

Whale optimization Algorithm is a metaheuristic algorithm which solves optimization problems by mimicking biological or physical phenomena. It worked in the bubble-net hunting strategy which is actually followed by the humpback whales. Results of WOA proved that it's a very much competitive algorithm than the other algorithms such as state-of-art-meta-heuristic and other typical methods [8].

There are three mathematical methods that are used by the algorithm to alternate between exploitation and exploration phases. They are described below:

1) Encircling Prey: Encircling prey is the phenomenon in which the humpback whale locates and surrounds the prey and the current best whale assumed to be closer to it and guides other whales to update their position towards the prey. The mathematical form of this phenomenon is given below:

$$\vec{D} = |\vec{C} \cdot \overrightarrow{X'}(t) - \vec{X}(t)| \qquad (3.1)$$

$$\vec{X}(t+1) = \overrightarrow{X'}(t) - \vec{A}.\vec{D}| \qquad (3.2)$$

$$\vec{A} = 2\vec{a} \cdot \vec{r} - \vec{a}| \qquad (3.3)$$

$$\vec{C} = 2 \cdot \vec{r} \qquad (3.4)$$

Where

$\overrightarrow{X'}$ is the position vector of the best solution,

T indicates the current location,

$\vec{X}$ refers to the current position,

$\vec{A}$ and $\vec{C}$ refers to the coefficient vector,

$\vec{a}$ is linearly decreased from 2 to 0

$\vec{r}$ refers to a random vector.

$\vec{D}$ refers to the distance between the best and current position

2) Attacking Prey: It’s an exploitation phase where bubble-net hunting strategy is followed to refine a solution. In real life humpback whales use this strategy to hunt their food such as fish. They create a bubble net to encircle their prey and swim in a 9-shaped spiral path while blowing bubbles and forms a net which helps them to trap the prey near the surface. This strategy is mathematically simulated by two mechanisms. They are given below:

   i. Shrinking Encircling Mechanism: It forces whales to approach the prey when $\vec{A}$<1

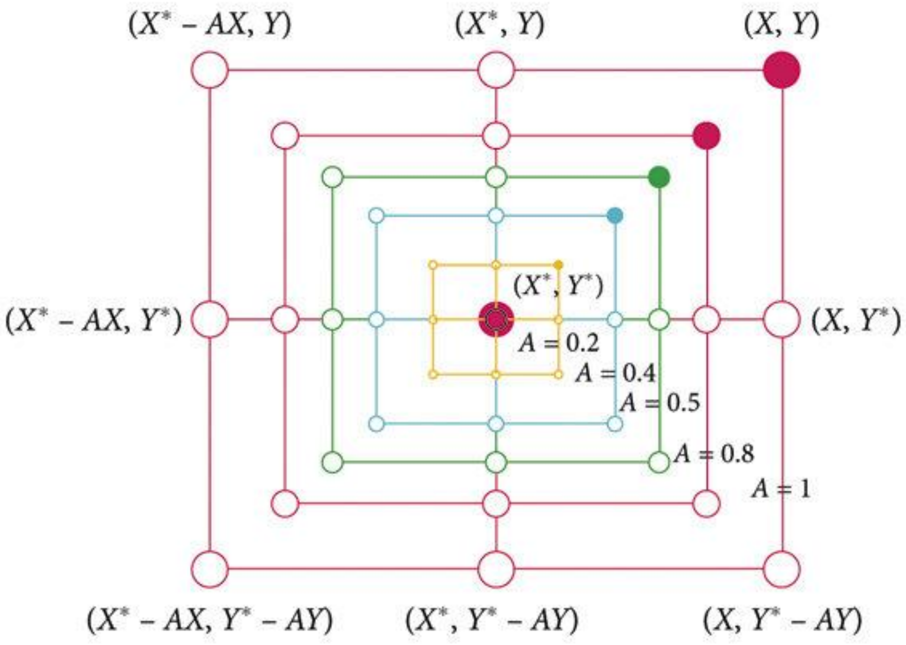

Figure 3.2: Shrinking encircling mechanism [42]

ii. Spiral Updating Position: Whale swims around the prey and update position using a logarithmic spiral. The equation is given below:

$$\vec{X}(t+1) = \overrightarrow{D'} \cdot e^{bl} \cdot \cos(2\pi l) + \overrightarrow{X'}(t) \qquad (3.5)$$

$$\overrightarrow{D'} = |\overrightarrow{X'}(t) - \vec{X}(t)| \qquad (3.6)$$

$$\vec{X}(t+1) = \begin{cases} \overrightarrow{|X'}(t) - \vec{A}.\vec{D}|f(x) & if\ p < 0.5 \\ \overrightarrow{D'} \cdot e^{bl} \cdot \cos(2\pi l) + \overrightarrow{X'}(t) & if\ p \geq 0.5 \end{cases} \qquad (3.7)$$

Where,

$\vec{X}(t+1)$ is the updated position of whale at the next next iteration (t+1),

$\overrightarrow{D'}$ refers to the distance between the ith whale and the best solution found so far.,

b refers to the logarithmic spiral shape,

l refers to a random number,

p determines whether the whale performs shrinking encircling behavior or spiral movement around the prey,

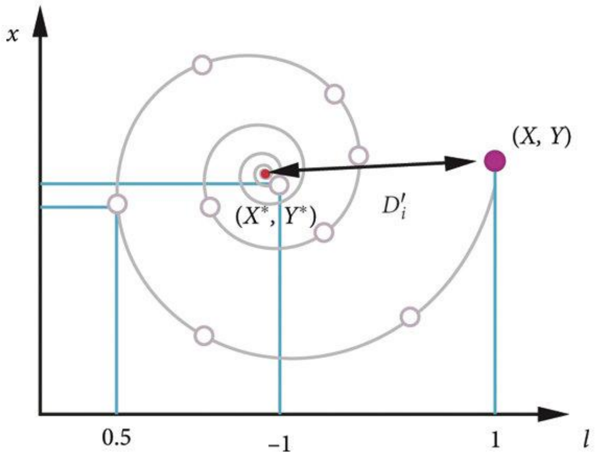

Figure 3.3: WOA spiral updating position **[42]**

iii. Searching Prey: It is the exploration phase which involves a global search. Whales enhance exploration and solves the stucking issue in a local optimum by moving randomly around other whales based on their positions [43].

The mathematical model is:

$$\vec{D} = |\vec{C} \cdot \vec{X}_{rand} - \vec{X}| \qquad (3.8)$$

$$\vec{X}(t+1) = \vec{X}_{rand} - \vec{X} \cdot \vec{D} \qquad (3.9)$$

Where $\vec{X}_{rand}$ is the random position vector chosen from the current population

3.5.1.1 Feature Selection Framework of WOA

Binary feature selection framework was made to use in WOA for feature selection. Here each whale is considered as a possible subset of features. Generally, WOA operates in a continuous search space, and it generates continuous values for feature selection. To make WOA enable to find or explore different combinations of feature subsets a threshold mechanism is used where the continuous whale position vectors are represented as a feature-mask vector. The binary decisions are Values> 0.5→ Feature Selected and Values<= 0.5→ Feature Not Selected.

3.5.1.2 Fitness Function Design

Fitness function is used to determine how good a solution is for a certain problem which was solved. Whales are guided to the best prey or solution by determining the smallest or highest value for that solution. The function uses calculations of encirclement, bubble-net hunting and random search for finding the optimal solution or feature subset in our case. In our study we designed the fitness function with three components. They are described below:

1. Prediction Error: A lightweight Gradient Boosting classifier was used to evaluate each whale or feature subset. The prediction error was computed using validation accuracy and good solution holds a lower value.

2. We set a penalty to the WOA to prevent it from selecting large feature subsets. A target number for selected subset is set as penalty and when whale selects features larger than the target the penalty becomes active.

   Penalty= 0.1*max (0, $N_{selected} - N_{target}$) (3.10)

   Where

   $N_{selected}$= selected feature's quantity by the current whale

   $N_{target=}$desired number of features
3. Stability Bonus: By obtaining frequency of each feature from previous WOA runs we added a stability aware bonus so that the high frequency features got priority in the upcoming runs.

Total fitness was counted by summing the above three components that are,

Fitness= BaseFitness+Penalty+StabilityBonus (3.11)

3.5.2 Stability Measurement Using Jaccard Similarity

Jaccard similarity is a method to measure the similarity between two sets. It is simple in nature and easy to implement. Consequently, it is applied in various domains of science and technology and found to be performed adequately. It is the ratio of the intersection size to the union size of the sample sets. It is widely used for visualization, classification and modeling and applied to compare vectors [44], [45]. We used Jaccard similarity in our study to ensure feature stability. Two approaches were followed to quantify the stability of each feature. They are:

1. The frequency of each feature was calculated using the formula:

   Frequency(f)= Number of runs where f was selected/Total runs

   Probabilistic importance of each feature and repeatedly selected biomarkers was determined from this approach.
2. The second approach was to use Jaccard similarity to measure the overlap between two subsets. It uses the formula:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (3.12)$$

Where

$|A \cap B|$= common selected features

$|A \cup B|$= number of total unique features across both subsets.

The overall stability score is determined by

$S = mean(J_1, J_2, \dots, J_k)$

The larger value of S ensures the reliability of the features for ovarian cancer classification.

Using all these approaches a final set of features were obtained for further research. We perform union operation among the set of best stable selected feature subsets and obtained a final set of 20 features.

3.6 Model Training

3.6.1 XGBoost Algorithm

We used Extreme Gradient Boosting or XGBoost for our study. It is based on the boosting principle and advance gradient boosting is implemented in the model. This classifier has enhanced predictive performance along with good efficiency and scalability. A minimization of differential loss function improves decision trees sequentially which is done by using gradient descent optimization [46]

Precise gradient and Hessian calculations are made by XGBoost classifier using second-order Taylor expansion of the loss function for more accurate weight update. [47]. XGBoost prevent overfitting by using different regularization techniques such as L1/L2 penalties, shrinkage, subsampling). These techniques required to perform to when the dataset is Small and high dimensional medical dataset. The classifier supports parallel processing, efficient tree pruning and fast convergence which reduces the training time.

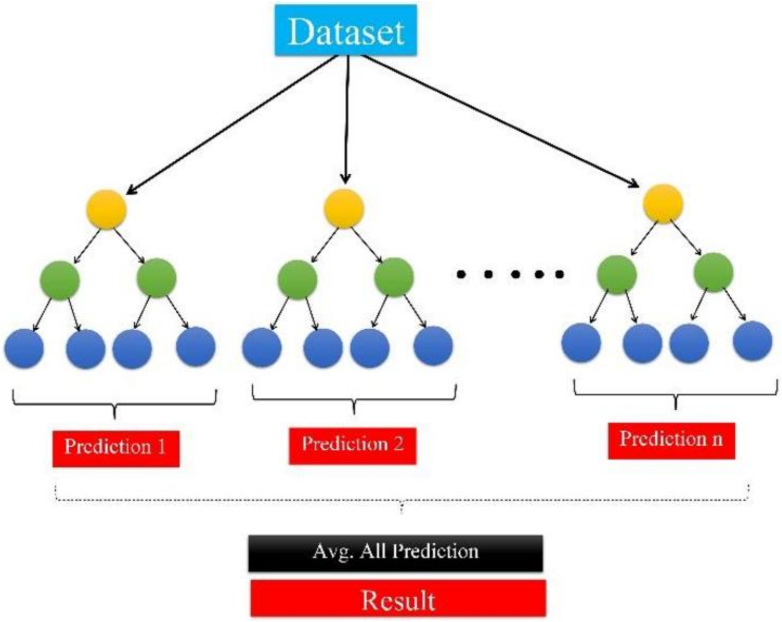

Figure 3.4: Schematic view of XGBoost model [48]

The above figure shows the schematic view of XGBoost classifier. The small trees are decision makers. Each tree makes a small prediction by looking at the data. If one tree makes a prediction the other tree looks at the error in the first tree's prediction. It then tries to correct the error by giving a more accurate prediction than the previous one. This process continues fixing the errors of previous trees. After all the trees made their prediction, the final prediction comes from the summation or combination of each prediction. This approach leads to predicting more accurate results as the result is made by many learning trees.

3.6.2 Hyperparameter Tuning

A key step in the development of machine learning algorithms is hypreparameter tuning. Especially complex classifiers like XGBoost use this to make results more reliable. Bias and variance have direct impact of hyperparameter tuning as it affects the architecture, depth, learning rate, regularization strength and sampling behavior of a model. Usually data size and model complexity can improve model generalization but hyperparameter tuning can do this even more [49]. Despite of class imbalance and high dimensionality of medical data hyperparamet.er tuning avoids the suboptimal minima and improves evaluation metrics on unseen medical data [50].

Hyperparameter tuning is extremely important in applications such as cancer prediction and precision medicine. Improved machine learning models may significantly enhance biomarker based prediction accuracy by minimizing overfitting and changing the model behavior to follow the biological signal patterns [51]. For the reasons stated above we also used hyperparameter tuning in our selected XGBoost classifier. Whale Optimization Algorithm was used to tune the hyperparameters to obtain optimal performance in ovarian cancer classification. The hyperparameters which we tuned were max_depth, learning_rate, n_estimators, subsample, colsample_bytree, min_child_weight, and gamma.

Each decision tree's complexity is controlled by max_depth. In small biomedical datasets, more interactions are caught by deeper trees when it memorizes noise. Best generalization can be made if moderate depth (5-6) is used [52]. We choose max_depth of 7 which ensures that the model can learn higher order biomarker interactions without high complexity.

n_estimators provides total number of boosting rounds. When there are more trees, overfitting is likely to occur but enhance the total performance. XGBoost is benefited by relatively larger number of tree when the learning rate is low [53]. Almost 498 trees were discovered by WOA to the perfect balance in our study. This number is large enough to recognize nonlinear relationships in cancer biomarker patterns while avoiding overfitting.

Each tree's contribution in the final model is measured by the learning rate. Supplementary boosting rounds combined with reduced learning rate enhance generalization [54]. We used learning_rate of 0.7 which enables the steady, continuous development of small biomarker patterns while avoiding rapid parameter change, which is in line with best practices in biomedical classification.

Subsampling helps to reduce overfitting by enabling tree diversity [55]. It helps to achieve good result by controlling the part of training samples which are used in building trees. We used subsample ratio of 0.96 which indicates the usefulness of the dataset. High dimensional cancer biomarker datasets have benefits of the little sampling noise as it improves generalization.

We used min_child_weights of 2.62 selected by WOA. This parameter helps to avoid the creation of leaves that are overly small and also prevents noisy sample groups. Clinical datasets achieve good result when the min_chil_weight is in the range of 1-5.

Gamma sets a threshold point for enabling tree splits. As a split needs a higher reduction in loss, a higher gamma leads the model to be more conservative. This helps to avoid unnecessary branching. Gamma is particularly critical because noisy features create misleading divisions in cancer classification. Tuning gamma can highly increase the prediction accuracy. In our study the value of gamma was chosen of 0.76 which reduces the likelihood of catching noise.

We choose hyperparameters that are stable and biologically meaningful which have fewer standard deviations. This helps to achieve the optimized result of predicting ovarian cancer significantly.

3.7 Performance Analysis Metrics

We used the general performance metrics like accuracy, precision, recall, F1-score, confusion matrix and Area Under the Curve (AUC) to evaluate our model. They are stated below:

Accuracy: It is used to measure the correctness of prediction. It done by calculating the ratio of summation of True Positives and True Negatives to the summation of True Positives, True Negatives, False Positives and False Negatives. It is the measure of among all the predictions how many predictions were correctly given by the model [56]. The formula for calculating accuracy is given by:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (3.13)$$

Precision: This reliability of the positive predictions of a model are evaluated by this metric which also known as positive predictive value [57]. It is calculated by the ratio of True Positives to the summation of all positive predictions. The formula is given by:

$$Precision = \frac{TP}{TP+FP} \qquad (3.14)$$

Recall: To determine how a model is capable to recognize all relevant cases in a dataset recall is used [57]. It is measured by the ratio of the true positives that the model predicted correctly to the number of all actual positive cases which includes the cases model did not predict as positives.

$$Precision = \frac{TP}{TP+FN} \qquad (3.15)$$

F1-Score: It is the combination of precision and recall which gives the harmonic mean of them. Model performance is balanced by this metric. When the recall and precision are of equal importance it comes to an action.

$$F1\,Score = \frac{2\times Precision\times Recall}{Precision+Recall} \qquad (3.16)$$

Confusion Matrix: A confusion matrix represents the predicted and actual classification which are given by a n x n matrix, where n is the number of different classes [58]. It shows the quantities of TP, TN, FP and FN which elaborates the model prediction in detail. The model classifies the misclassification as well as highlights errors made by the classifier.

ROC and AUC: The representation of the performance of a classifier across different cutoff values using a curve is known as Receiver Operating Characteristics curve or ROC. The Area Under the curve (AUC) gives a measure of the model’s differentiating capability between classes. The greater the value of the AUC the better the predictive accuracy of the model on different thresholds.

3.8 Conclusion

The chapter gives an in-detail explanation of the materials and methods we used in our study.

We collected an ovarian cancer dataset from Mendeley Data platform as the first step. The dataset has some missing values, so we have to preprocess it. We filled the missing values with the meaning of the corresponding column. We do not perform any categorical encoding as the dataset contains only numerical values. These steps made the dataset suitable for the steps.

We then evaluate three different models using the processed dataset. It helped us to make a decision which classifier to choose. We chose XGBoost classifier for further work based on its output metrics.

Our goal was to reduce the features as much as possible while maintaining the model's reliable predictive capability. We used Whale Optimization Algorithm for feature selection. We did not select features randomly rather based on their stability across different runs. Jaccard similarity and the frequency measure was used to assign frequencies of each feature and select the most stable feature for the final feature set. This helped us to recognize the most important and stable features for ovarian cancer prediction.

After obtaining the final reduced feature set we used XGBoost classifier to train our model. It's a powerful gradient boosting algorithm which is scalable and shows good performance in handling structured data. To obtain better results we tuned the hyperparameters of the XGBoost classifier including max_depth, learning_rate, n_estimators, subsample, gamma and min_child_weight. This fine tuning balanced the complexity and ensured generalization of the model. We then evaluated the model using various performance metrics such as accuracy, precision, recall, F1-score, and AUC-ROC which actually represents the model's capability of giving reliable results.

This chapter is the reflection of the total work we did to predict ovarian cancer accurately. The next chapter will present the result of our study along with the effectiveness of it in predicting ovarian cancer.

Chapter 4: Results and Discussion

4.1 Introduction

The result of this research gives a detailed evaluation of the performance of different hyperparameter tuned and untuned machine learning models on all features as well as a set of stable features selected by the metaheuristic whale optimization algorithm in predicting ovarian cancer. Figure 1 shows the overview of the training process of our study.

This chapter demonstrates all the outcomes of our study. Our selected dataset had a total of 349 samples with 49 features.

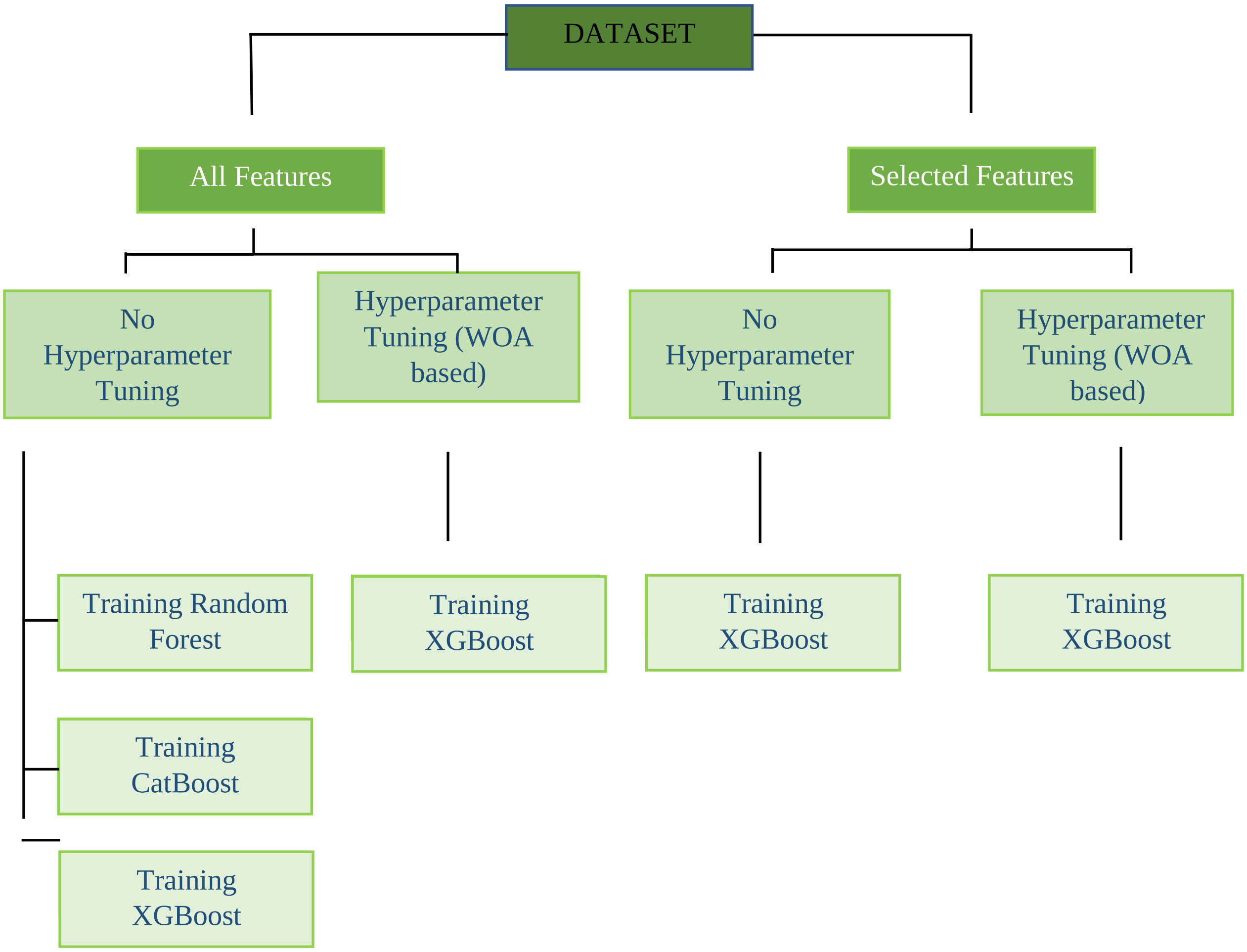

Figure 4.1: Visualization of model training process

4.2 Performances of Random Forest, XGBoost and CatBoost Hyperparameter Untuned Classifier with All Features

The first classifier used in our research was Random Forest. It was trained with untuned hyperparameters with all features. It achieved accuracy of 89.69%, 90.28% precision, 89.69% recall, 89.64% F1- score and 94.58% AUC.

We used another classifier called CatBoost for better performance. It achieved accuracy of 87.11%, 88.13% precision, 87.11% recall, 87.01% F1- score and 95.95% AUC.

We also used XGBoost classifier which achieved accuracy of 89.69%, 90.09% precision which indicates that the model predicts ovarian cancer 90.09% of the time correctly and minimizes the false positives. It achieved 89.69% recall which suggests that the model successfully predicts 89.69% of the actual cancer patients. The 89.65% F1-score balances the precision and recall, and the 94.85% AUC indicates that the model is highly effective at differentiate between positive and negative classes across all possible classification thresholds.

Table 4.1: Performance of three different models with all features

Hyperparameter	**Model**	**Accuracy**	**F1**	**Precision**	**Recall**	**AUC**
Untuned	**Random Forest**	0.8969	0.8964	0.9028	0.8969	0.9458
	CatBoost	0.8711	0.8701	0.8813	0.8711	0.9595
	XGBoost	0.8969	0.8965	0.9009	0.8969	0.9485

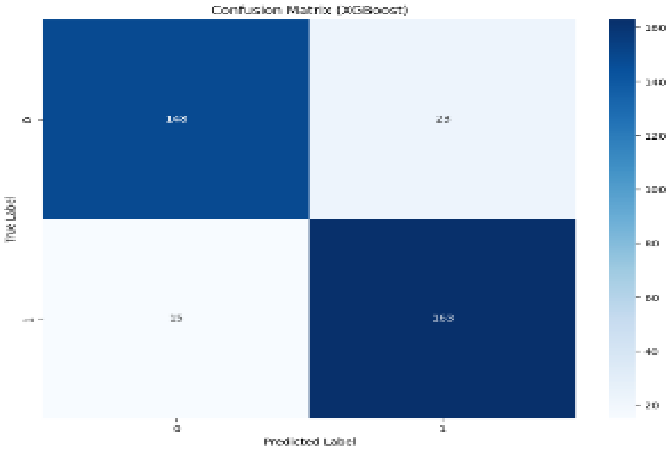

Figure 4.2: Confusion matrix of untuned XGBoost classifier with all feature

The above figure shows the confusion matrix for the XGBoost classifier with untuned hyperparameters. It correctly predicted 148 negatives and 163 positives with 23 false positive and 15 false negatives. The finding ensures that the model has good ability to predict the true negative and true positives, but the model can be further optimized to minimize the misclassification.

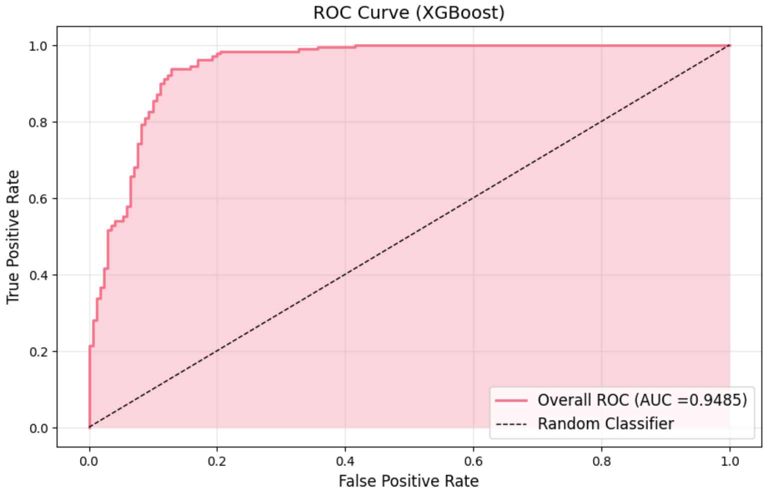

Figure 4.3: ROC curve for hyperparameter untuned XGBoost classifier

The ROC curve is a graph that shows the performance of the binary classifier across different thresholds. It illustrates how well a model discriminate across classes by graphing the True Positive Rate (Sensitivity) versus the False Positive Rate (1-Specificity).

The ROC curve of the XGBoost classifier shows an upright rise which closely follows the top-left corner which indicates that the classifier has high sensitivity with low false positives. The Area Under the Curve (AUC) has a value of 0.9485 which signifies that the model has near perfect classification capability and can identify 94.85% cases correctly. This evaluation value makes the model highly reliable for real-world medical applications.

4.3 Performance of WOA based Hyperparameter Tuned XGBoost Classifier by with All Features

After the untuned XGBoost model evaluation we have performed hyperparameter tuning in the model using Whale Optimization Algorithm which is a metaheuristic algorithm. Meta-heuristic algorithms which are nature inspired follows biological or physical incidents to solve the optimization problems. [59]. The algorithm solves complex optimization problems by the bubble-net hunting behavior of humpback whales. The hyperparameter tuned model using WOA achieved an accuracy of about 90.27%, 89.97% recall and AUC of about 96.59%. Table 4.2 shows the performance metrics of the model.

Table 4.2: Performance of hyperparameter tuned XGBoost classifier using WOA with all features

Hyperparameter	**Model**	**Accuracy**	**F1**	**Precision**	**Recall**	**AUC**
Tuned	**XGBoost**	0.9027	0.9025	0.9049	0.9027	0.9544

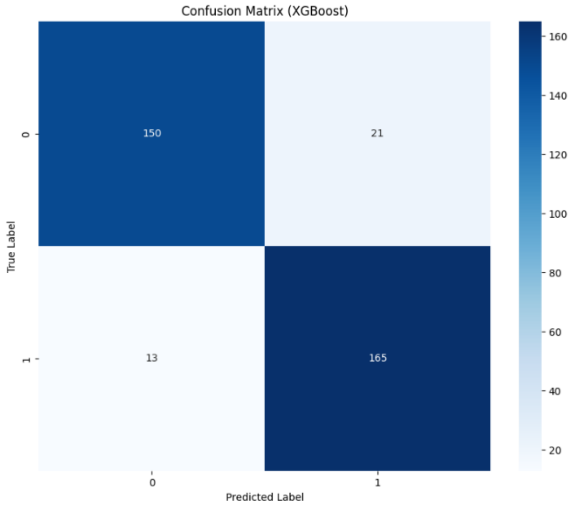

Figure 4.4: Confusion matrix of WOA based hyperparameter tuned XGBoost classifier with all features

The figure above is the confusion matrix of the WOA based hyperparameter tuned XGBoost model. It shows that the model correctly predicted 150 negatives and 165 positives with only 21 false positive and 13 false negative. The specificity is about 87.72% which implies that the model can effectively distinguish true negatives and demonstrates overall strong classification capability with less error.

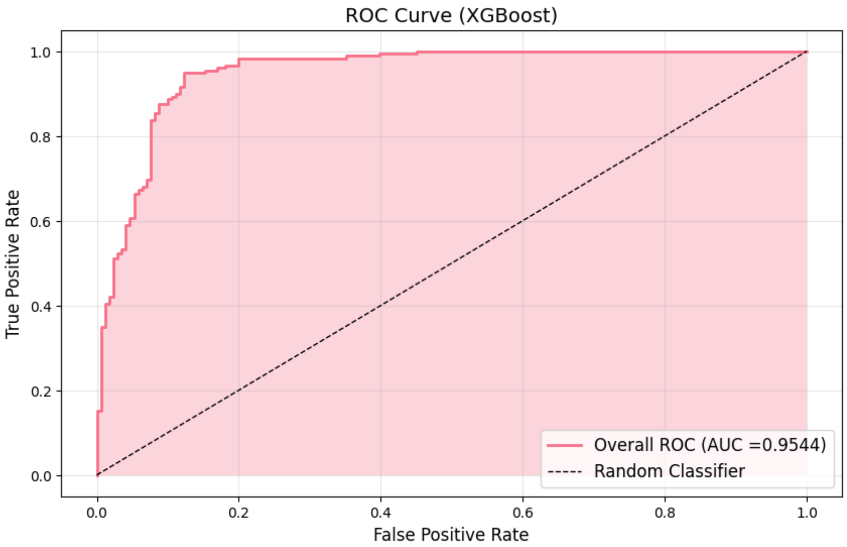

Figure 4.5: ROC curve for WOA based hyperparameter tuned XGBoost classifier

The above figure is the ROC curve of the WOA based hyperparameter tuned XGBoost model.

The Area Under the Curve (AUC) which is derived from the ROC curve is a comprehensive metric that gives the overall discriminatory power of the model. The model achieves about 95.44% of AUC which indicates the model has a better

ability to classify instances across all thresholds correctly. This reflects the tuned XGBoost model has a good classification performance. ROC and AUC are indispensable tools for selecting and comparing models in domains where misclassification costs vary significantly, such as healthcare, finance, and security [60]

4.4 Performance of Hyperparameter Untuned XGBoost Classifier with Selected Features

Feature selection method which is a reduction technique finds and pick a small feature subset by removing unnecessary, redundant or noisy features from the original feature set. Better learning performance, such as increased learning accuracy, reduced computing cost, and improved model interpretability, can typically result from feature selection [61]. We used the Whale Optimization Algorithm for feature selection. WOA selected the most stable set of 20 features among the 49 features. The selected features are HE4, DBIL, CEA, MPV, NEU, CA125, HCT, LYM%, ALP, PLT, Age, PCT, BASO#, BUN, BASO%, IBIL, MONO%, GLO, Ca, CEA. The untuned XGBoost model was used to evaluate the metrics with these selected features. Table 4.3 shows the evaluation metrics.

Table 4.3: Performance of untuned XGboost classifier with selected features

Method	**Hyperparameter**	**Model**	**Accuracy**	**F1**	**Precision**	**Recall**	**AUC**
WOA	Untuned	XGBoost	0.8769	0.8765	0.8801	0.8769	0.9528

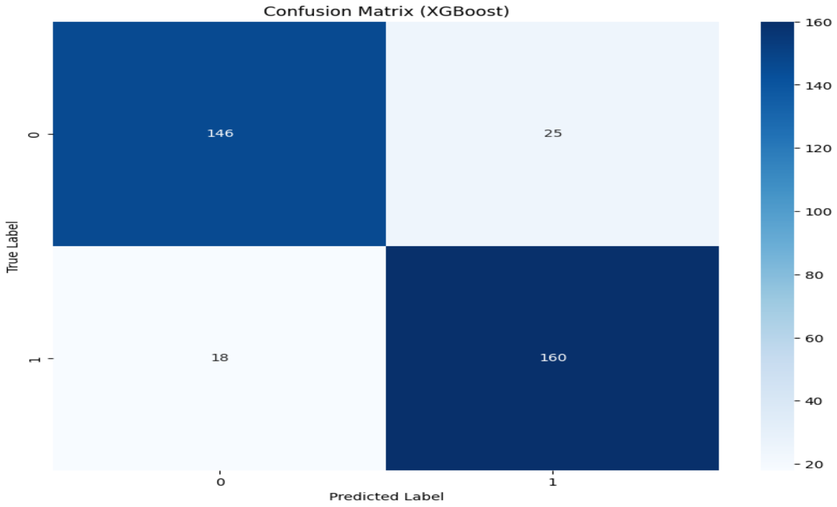

Figure 4.6: Confusion matrix of untuned XGBoost classifier with selected features

The figure above is the confusion matrix of the untuned XGBoost model. It demonstrates that the model correctly predicted 146 negatives and 160 positives with only 25 false positive and 18 false negative. This means that the model correctly identified 146 cases as non-cancer which means that this patient has no cancer and the model did not classify them as cancer. On the other hand, 160 cases have cancer and the model also predicted this correctly. This leads to the precision to hold a value of about 88.01%. But the model misclassified some cases as cancer while they are not and non-cancer while it is cancer.

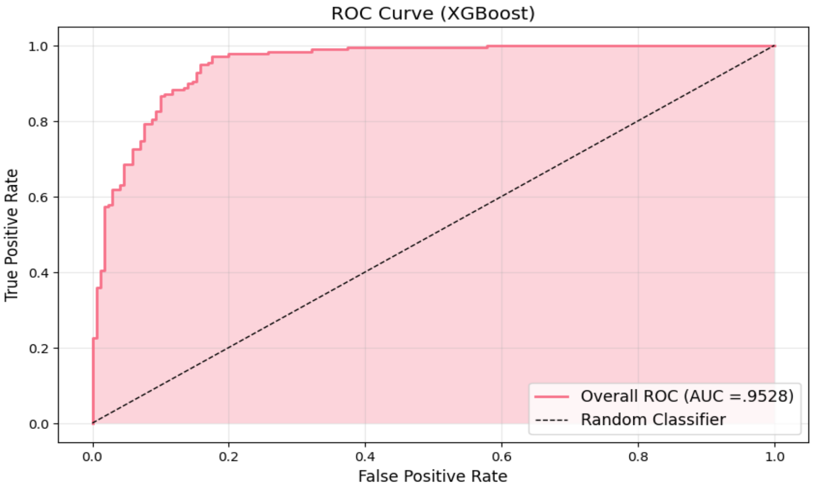

Figure 4.7: ROC curve for untuned XGBoost classifier

The figures show the ROC curve of the model which demonstrates the discrimination ability of the model. The value of AUC is about 0.9528 or 95.28% which specifies that the classifier has a decent balance of sensitivity and specificity. The curve is tending to near to the upper left corner which specifies that the model has relatively a smaller number of false positives and false negatives which makes the model reliable.

4.5 Performance of WOA based Hyperparameter Tuned XGBoost Classifier with Selected Features

The model was then tested with WOA based hyperparamters with the selected features.

It shows that the model has improved in all the metrics. The accuracy was 89.69%, precision, recall and AUC were 90.04%, 89.69% and 95.36% respectively.

Table 4.4: Performance of hyperparameter tuned XGboost classifier using WOA with selected features

Method	**Hyperparameter**	**Model**	**Accuracy**	**F1**	**Precision**	**Recall**	**AUC**
WOA	Tuned	XGBoost	0.8969	0.8967	0.9004	0.8969	0.9536

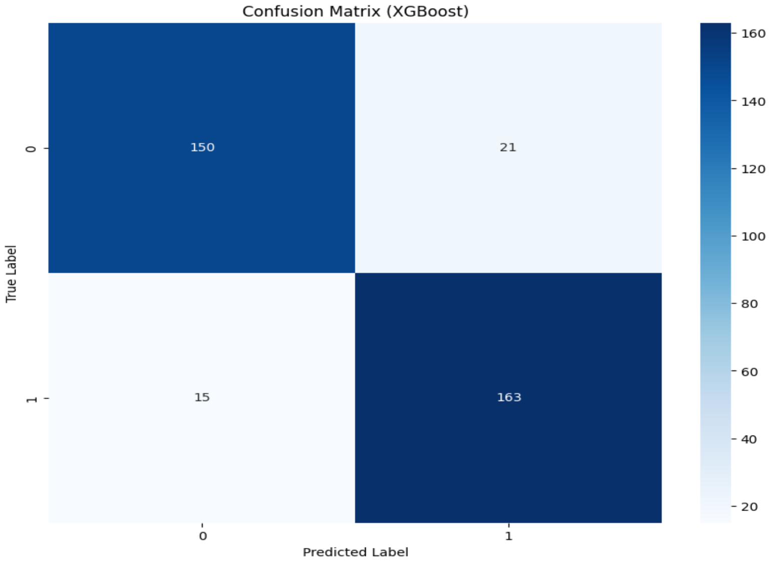

Figure 4.8: Confusion matrix of WOA based hyperparameter tuned XGBoost classifier with selected features

Figure 4.6 shows the confusion matrix of WOA based hyperparameter tuned XGBoost classifier with selected features. In the matrix 150 cases were correctly predicted of the negative class and 163 cases of the positive class by the model. The high number of the True Positives and True Negatives are the indication of the correctly identifying capability of both classes of the model. However, the model misclassified 21 cases as positive while they were negative and 15 cases as negative while, they were positive. So, the model can be further optimized to minimize this misclassification.

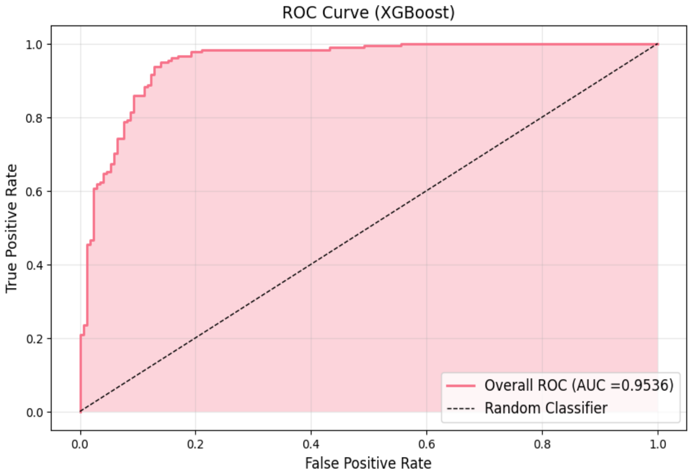

Figure 4.9: ROC curve of WOA based hyperparameter tuned XGBoost classifier with selected features

The above figure shows the ROC curve of the tuned XGBoost model with all features. The pink line is near to the top left-hand corner of the curve. This reflects the high classification competence of the model. The AUC is 95.36% which indicates that the model is near to perfection and has a good correctly classifying ability of the classes.

4.6 Impact and Importance of features:

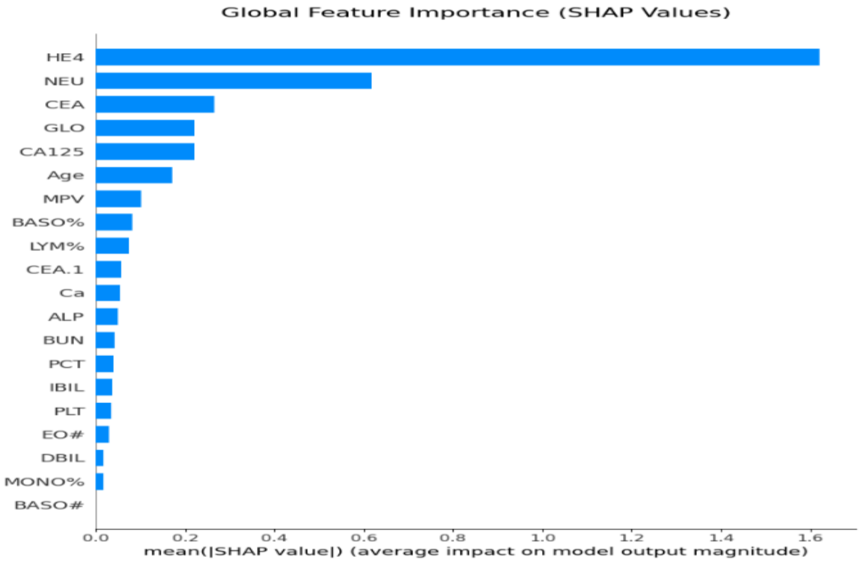

Figure 4.10: Feature Importance Bar Chart Using WOA for Feature Selection and and hyperparameter tuned XGBoost for Modeling.

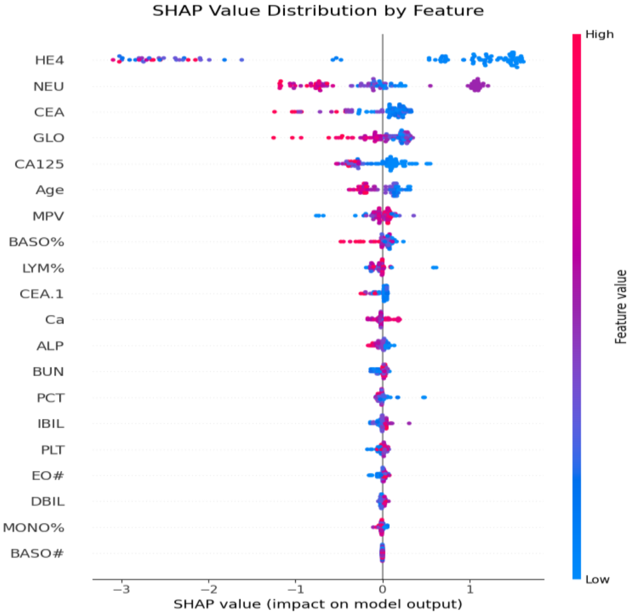

Figure 4.11. Summary Plot Using WOA for Feature Selection and hyperparameter tuned XGBoost for Modeling.

The above figures show the graphs of SHAP value analysis. Using SHAP values, a model addictive explanation technique, each prediction is explained by the contribution of the dataset's features to the model's output. More precisely, SHAP approximates Shapley values, a notion from game theory that addresses the problem of figuring out how each subset of features contributes to a model's prediction given a dataset of m features..[62]

The graphs demonstrate the influence and relative significance of the selected features in predicting ovarian cancer using SHAP value analysis. The result sorts out the most important features in descending order from all the selected features. It indicates that among all selected features, the biomarker HE4, NEU and CEA shows the great impact on the model’s prediction.

4.7 Model Performance Comparison

Table 4.5: Model performance comparison between three different models

Hyperparameter	Model	Accuracy	F1	Precision	Recall	AUC
Untuned	**Random Forest**	0.8969	0.8964	0.9028	0.8969	0.9458
	CatBoost	0.8711	0.8701	0.8813	0.8711	0.9595
	XGBoost	0.8969	0.8965	0.9009	0.8969	0.9485

From Table 4.3 we can see the performance metrics of the three untuned underline baseline models- Random Forest, CatBoost and XGBoost models. Among the models Random Forest and XGBoost classifiers have identical accuracy (0.8969) and F1 score (~0.8965). The highest precision was achieved by Random Forest model (0.9028) which suggests that it predicts fewer false positives than other two algorithms. CatBoost classifier achieved lower accuracy (0.8711) and F1 score (0.8701) than the others but achieved the highest AUC (0.9595) among all the models. It indicates that catboost is more effective in selecting positive and negative classed than other two models. The XGBoost classifier achieved accuracy and recall almost similar to the Random Forest classifier but achieved slightly lower precision (0.9009) and AUC (0.9485). This indicates that it predicts the negative cases incorrectly than the Random Forest classifier.

Table 4.6: Model performance comparison between different variants of XGBoost classifier

Features	Hyperparameter	Model	Accuracy	F1	Precision	Recall	AUC
All	**Untuned**	**XGBoost**	0.8969	0.8965	0.9009	0.8969	0.9485
	Tuned	**XGBoost**	0.9027	0.9025	0.9049	0.9027	0.9544
Selected	**Untuned**	XGBoost	0.8769	0.8765	0.8801	0.8769	0.9528
	Tuned	XGBoost	0.8969	0.8967	0.9004	0.8969	0.9536

XGBoost classifiers was selected for further research because the metrics it produced have the most balanced performance across all the other classifiers. Also, it offers great flexibility for optimization and strong interpretability features which are crucial for medical research. So based on feature selection and hyperparameter optimization Table 4.4 shows the different results of the

XGBoost classifier. Among all the variants tuned XGBoost, with all features obtained the highest accuracy of 90.27% and an AUC of 95.44%. It indicates that performance of a model can be substantially improving by hyperparameter tuning. Performance was slightly decreased when the model was trained with the features selected by the Whale Optimization Algorithm. It indicates that some of the features holding good predictive information were not present in the selected features. However, the tuned model again performs significantly well compared to the untuned model with selected features demonstrating that the hyperparameter optimization has great effectiveness regarding models' performance.

4.8 Comparison with Previous Works

Table 4.7: Model performance comparison with previous works

Work	Dataset Used	Method used for Feature Selection/Importance	Classifier Used	Accuracy	F1-score	Precision	Recall	AUC
[63]	[64]	Boruta	CatBoost	0.8952	0.8945	0.9073	0.8952	0.9502
[65]	[64]	Information gain indices and the Gini index.	RF	0.8675	0.8801	-	0.9160	0.9250
[66]	[64]	Student's t-test and the Mann–Whitney U-test	RF, GBM	0.88	0.89	-	0.95	0.87
Ours	[64]	Whale Optimization Algorithm	XGBoost	0.8969	0.8967	0.9004	0.8969	0.9536

The Table 4.5 shows the comparison of our proposed model with previously models such as CatBoost, Random Forest, Gradient Boosting Machine (GBM) evaluated by different researchers on the same dataset in ovarian cancer prediction. The research performed by Tu˘gçe Öznacar et al., used CatBoost along with Boruta feature selection method and obtained an accuracy of 89.52% and an AUC of 95.02% with selected 20 features [63]. Another study perfromed by Seyed Mohammad Ayyoubzadeh et al., used information gain indices and Gini index for determining the feature importance with Random Forest model and achieved a lower accuracy of 86.75% and an AUC of 92.5% [67]. A third study conducted by Md. Martuza Ahamad et al., employed statistical significant tests (Student's t-test and Mann-Whitney U-test) combined with Random Forest and

Gradient Boosting Machine (GBM) and achieved an accuracy of 88% and an AUC of 87%[66]. In contrast our model outperformed all previous results in terms of AUC and accuracy. We used Whale Optimization Algorithm and Jaccard Similarity for stable feature selection combined with XGBoost classifier and obtained an accuracy of 89.69% and an AUC of 95.36% with only selected 20 features. These improvements specifies that the Whale Optimization Algorithm is highly effective in selecting informative and stable features. It also demonstrates that XGBoost classifier has superior predictive capability in predicting ovarian cancer classification.

4.9 Conclusion

A detailed evaluation of several machine learning models and feature selection methods applied to the ovarian cancer dataset was presented by the chapter. The evaluation began by showing a comparison between three baseline classifiers- Random Forest, CatBoost and XGBoost with all features. Among them XGBoost was most stable and showed competitive performance across all metrics. Hence, it justified the selection of XGBoost for further investigation. Performance of XGBoost was analyzed by both hyperparameter tuning and feature optimization. The Whale Optimization Algorithm identified the most stable feature subset, and the model was evaluated on them both before and after hyperparameter tuning. The result gives valuable insights such as feature selection improved model efficiency, hyperparameter tuning consistently improved model's accuracy and generalization. In comparison with the previous works the proposed method achieved either comparable or superior performance across most of the reported metrics. In conclusion we can say that the overall findings ensures that the combination of WOA based feature selection method with XGBoost classifier, and proper hyperparamter tuning can serve as an effective tool for predicting ovarian cancer

Chapter 5: Conclusion and Future Work

5.1 Conclusion

Ovarian cancer is one of the most common types of cancer in women. The survival rates are much lower that other cancers that affect women. In ranking of cancer-related death among women it ranks fifth. So, it is very important to detect ovarian cancer at the early stage although it is very difficult because symptoms often develop until later stages. The objective of our study was i) investigate the ovarian cancer dataset and sort out the important and stable features from a large number of features for ovarian cancer prediction, ii) to optimize the XGBoost classifier for using Whale Optimization Algorithm to obtain reliable outcomes, iii) to explore the performance of optimized XGBoost classifier for predicting the ovarian cancer, iv) to make a comparison between our study and the previous research. To achieve our goals, we conducted our study in several steps. At first, we collected an ethically approved effective dataset and investigated it. Then we used this dataset to evaluate various machine learning classifiers and selected the best one. To reduce the large feature overhead we have performed feature engineering using WOA which gives a stable feature set. In order to achieve higher evaluation metrics, we tuned the hyperparameters of the XGBoost classifier using the same WOA optimization method. Finally, we achieved the ovarian cancer prediction metrics which were then compared with prior research for validation. Our model achieved accuracy of 87.69% with selected features and hyperparameter untuned XGBoost classifier. Where the accuracy was improved to 89.69% when we perform hyperparameter tuning on the XGBoost classifier. Achieving this we found that our model outperformed other models which also used the the same dataset. This indicates that our model is more reliable than that of the prior models. From a healthcare perspective we hope that this improvement we have made will contribute significantly to the prediction of ovarian cancer.

5.2 Future Works

Although our research able to build a model that has achieved strong predictive performance with a selection of stable feature sets using Whale Optimization Algorithm and XGBoost classifier there remains some potential directions for further investigation. The main improvement should be the increment of datasets and addition of other clinical and molecular parameters. This can improve generalizability and dataset biasness. We used basic WOA for feature selection and hyperparameter tuning. Chaotic WOA [68] or Modified WOA (mWOA) can be used to reduce premature convergence and Hybrid WOA-PSO [69] or WOA-GA for enhanced exploration. These hybrid models can perform better feature selection ability and convergence speed than single optimization algorithm. Our current WOA framework contains only a single objective fitness function. Multi object frameworks can be added to optimize the model further. NSGA-II, MOPSO [70] can be used to balance various objectives.

By addressing the improvement suggestions given above the prediction of ovarian cancer could be revolutionize.

Refferences

[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, Apr. 2023, doi: 10.3389/fmed.2023.1145487.

[2] U. A. 1 Matulonis *et al.*, “Ovarian cancer (Primer),” 2016, doi: 10.1038/nrdp.2016.61.

[3] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2018,” *CA: A Cancer Journal for Clinicians*, vol. 68, no. 1, pp. 7–30, 2018, doi: 10.3322/caac.21442.

[4] M.-K. Hong and D.-C. Ding, “Early Diagnosis of Ovarian Cancer: A Comprehensive Review of the Advances, Challenges, and Future Directions,” *Diagnostics*, vol. 15, no. 4, p. 406, Jan. 2025, doi: 10.3390/diagnostics15040406.

[5] M. N. Wernick, Y. Yang, J. G. Brankov, G. Yourganov, and S. C. Strother, “Machine Learning in Medical Imaging,” *IEEE Signal Processing Magazine*, vol. 27, no. 4, pp. 25–38, July 2010, doi: 10.1109/MSP.2010.936730.

[6] T. Öznacar and T. Güler, “Prediction of Early Diagnosis in Ovarian Cancer Patients Using Machine Learning Approaches with Boruta and Advanced Feature Selection,” *Life*, vol. 15, no. 4, p. 594, Apr. 2025, doi: 10.3390/life15040594.

[7] “Whale optimization algorithm and its application in machine learning,” in *Handbook of Whale Optimization Algorithm*, Academic Press, 2024, pp. 69–80. doi: 10.1016/B978-0-32-395365-8.00011-7.

[8] S. Mirjalili and A. Lewis, “The Whale Optimization Algorithm,” *Advances in Engineering Software*, vol. 95, pp. 51–67, May 2016, doi: 10.1016/j.advengsoft.2016.01.008.

[9] L. Deng and S. Liu, “Deficiencies of the whale optimization algorithm and its validation method,” *Expert Systems with Applications*, vol. 237, p. 121544, Mar. 2024, doi: 10.1016/j.eswa.2023.121544.

[10] E. M. Wilson, R. N. Eskander, and P. S. Binder, “Recent Therapeutic Advances in Gynecologic Oncology: A Review,” *Cancers*, vol. 16, no. 4, p. 770, Jan. 2024, doi: 10.3390/cancers16040770.

[11] Y. Jiang, C. Wang, and S. Zhou, “Artificial intelligence-based risk stratification, accurate diagnosis and treatment prediction in gynecologic oncology,” *Seminars in Cancer Biology*, vol. 96, pp. 82–99, Nov. 2023, doi: 10.1016/j.semcancer.2023.09.005.

[12] O. Ozhan, Z. Kucukakcali, and I. B. Cicek, “Machine learning-based ovarian cancer prediction with XGboost and stochastic gradient boosting models.,” *Medicine Science*, vol. 12, no. 1, p. 231, Mar. 2023, doi: 10.5455/medscience.2022.09.207.

[13] “The impact of ovarian cancer on individuals and their caregivers: A qualitative analysis - Tan - 2021 - Psycho-Oncology - Wiley Online Library.” Accessed: Dec. 10, 2025. [Online]. Available: https://onlinelibrary.wiley.com/doi/epdf/10.1002/pon.5551

[14] I. J. Jacobs and U. Menon, “Progress and Challenges in Screening for Early Detection of Ovarian Cancer,” *Molecular & Cellular Proteomics*, vol. 3, no. 4, pp. 355–366, Apr. 2004, doi: 10.1074/mcp.R400006-MCP200.

[15] H. Snyder, “Literature review as a research methodology: An overview and guidelines,” *Journal of Business Research*, vol. 104, pp. 333–339, Nov. 2019, doi: 10.1016/j.jbusres.2019.07.039.

[16] G. Lame, “Systematic Literature Reviews: An Introduction,” *Proc. Int. Conf. Eng. Des.*, vol. 1, no. 1, pp. 1633–1642, July 2019, doi: 10.1017/dsi.2019.169.

[17] J. Davis, K. Mengersen, S. Bennett, and L. Mazerolle, “Viewing systematic reviews and meta-analysis in social research through different lenses,” *SpringerPlus*, vol. 3, no. 1, p. 511, Sept. 2014, doi: 10.1186/2193-1801-3-511.

[18] C. Kuzudisli, B. Bakir-Gungor, N. Bulut, B. Qaqish, and M. Yousef, “Review of feature selection approaches based on grouping of features,” *PeerJ*, vol. 11, p. e15666, July 2023, doi: 10.7717/peerj.15666.

[19] D. L. P and M. P, “Comparative Analysis of Machine Learning Frameworks for Robust Ovarian Cancer Detection Using Feature Selection and Data Balancing,” *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 16, no. 6, June 2025, doi: 10.14569/IJACSA.2025.0160687.

[20] J. Cai, Z.-J. Lee, Z. Lin, C.-H. Hsu, and Y. Lin, “An Integrated Algorithm with Feature Selection, Data Augmentation, and XGBoost for Ovarian Cancer.,” *Mathematics (2227-7390)*, vol. 12, no. 24, p. 4041, Dec. 2024, doi: 10.3390/math12244041.

[21] B. Wickramasinghe and G. Regisford, “A Comparative Study to Predict Ovarian Cancer,” *Journal of the South Carolina Academy of Science*, vol. 22, no. 2, Sept. 2024, [Online]. Available: https://scholarcommons.sc.edu/jscas/vol22/iss2/8

[22] L. Akter and N. Akhter, “Ovarian Cancer Prediction from Ovarian Cysts Based on TVUS Using Machine Learning Algorithms,” in *Proceedings of the International Conference on Big Data, IoT, and Machine Learning*, M. S. Arefin, M. S. Kaiser, A. Bandyopadhyay, Md. A. R. Ahad, and K. Ray, Eds., Singapore: Springer, 2022, pp. 51–61. doi: 10.1007/978-981-16-6636-0_5.

[23] E. Bartz, T. Bartz-Beielstein, M. Zaefferer, and O. Mersmann, Eds., *Hyperparameter Tuning for Machine and Deep Learning with R: A Practical Guide*. Springer Nature, 2023. doi: 10.1007/978-981-19-5170-1.

[24] R. Shetty, S. Gupta, V. Mediratta, S. Rai, and M. Geetha, “Optimizing Machine Learning-Based Ovarian Cancer Prediction Through Normalization Strategies,” *IEEE Access*, vol. 13, pp. 128974–128995, 2025, doi: 10.1109/ACCESS.2025.3590871.

[25] H. Dhingra and R. Shetty, “Comparative Study of Machine Learning and Deep Learning Models for Early Prediction of Ovarian Cancer,” *IEEE Access*, vol. 13, pp. 87336–87349, 2025, doi: 10.1109/ACCESS.2025.3567081.

[26] A. Sorayaie Azar *et al.*, “Application of machine learning techniques for predicting survival in ovarian cancer,” *BMC Med Inform Decis Mak*, vol. 22, no. 1, p. 345, Dec. 2022, doi: 10.1186/s12911-022-02087-y.

[27] Y. Sun and B. Wen, “Machine-learning diagnostic models for ovarian tumors,” *Heliyon*, vol. 10, no. 19, Oct. 2024, doi: 10.1016/j.heliyon.2024.e36994.

[28] C. Tang, T. Gao, Y. Li, and B. Chen, “EEG channel selection based on sequential backward floating search for motor imagery classification,” *Front. Neurosci.*, vol. 16, Oct. 2022, doi: 10.3389/fnins.2022.1045851.

[29] S. L. J. M and S. P, “Innovative approach towards early prediction of ovarian cancer: Machine learning- enabled XAI techniques,” *Heliyon*, vol. 10, no. 9, May 2024, doi: 10.1016/j.heliyon.2024.e29197.

[30] S. D. Patil, P. J. Deore, and V. B. Patil, “An Intelligent Computer Aided Diagnosis System for Classification of Ovarian Masses using Machine Learning Approach,” *International Research Journal of Multidisciplinary Technovation*, pp. 45–57, Apr. 2024, doi: 10.54392/irjmt2434.

[31] S. M. Ayyoubzadeh, M. Ahmadi, A. B. Yazdipour, F. Ghorbani-Bidkorpeh, and M. Ahmadi, “Prediction of ovarian cancer using artificial intelligence tools,” *Health Science Reports*, vol. 7, no. 7, p. e2203, 2024, doi: 10.1002/hsr2.2203.

[32] M. Lu *et al.*, “Using machine learning to predict ovarian cancer,” *International Journal of Medical Informatics*, vol. 141, p. 104195, Sept. 2020, doi: 10.1016/j.ijmedinf.2020.104195.

[33] C. Ding and H. Peng, “Minimum redundancy feature selection from microarray gene expression data,” *J. Bioinform. Comput. Biol.*, vol. 03, no. 02, pp. 185–205, Apr. 2005, doi: 10.1142/S0219720005001004.

[34] R. G. Moore *et al.*, “A novel multiple marker bioassay utilizing HE4 and CA125 for the prediction of ovarian cancer in patients with a pelvic mass,” *Gynecologic Oncology*, vol. 112, no. 1, pp. 40–46, Jan. 2009, doi: 10.1016/j.ygyno.2008.08.031.

[35] A. Arfiani and Z. Rustam, “Ovarian cancer data classification using bagging and random forest,” *AIP Conf. Proc.*, vol. 2168, no. 1, p. 020046, Nov. 2019, doi: 10.1063/1.5132473.

[36] A. El-Nabawy, N. El-Bendary, and N. A. Belal, “Epithelial Ovarian Cancer Stage Subtype Classification using Clinical and Gene Expression Integrative Approach,” *Procedia Computer Science*, vol. 131, pp. 23–30, Jan. 2018, doi: 10.1016/j.procs.2018.04.181.

[37] M. M. Ahamad *et al.*, “Early-Stage Detection of Ovarian Cancer Based on Clinical Data Using Machine Learning Approaches,” *Journal of Personalized Medicine*, vol. 12, no. 8, p. 1211, Aug. 2022, doi: 10.3390/jpm12081211.

[38] B. Bala and S. Behal, “A Brief Survey of Data Preprocessing in Machine Learning and Deep Learning Techniques,” in *2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Oct. 2024, pp. 1755–1762. doi: 10.1109/I-SMAC61858.2024.10714767.

[39] A. Parmar, R. Katariya, and V. Patel, “A Review on Random Forest: An Ensemble Classifier,” in *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*, vol. 26, J. Hemanth, X. Fernando, P. Lafata, and Z. Baig, Eds., in Lecture Notes on Data Engineering and Communications Technologies, vol. 26. , Cham: Springer International Publishing, 2019, pp. 758–763. doi: 10.1007/978-3-030-03146-6_86.

[40] J. R. Vergara and P. A. Estévez, “A review of feature selection methods based on mutual information,” *Neural Comput & Applic*, vol. 24, no. 1, pp. 175–186, Jan. 2014, doi: 10.1007/s00521-013-1368-0.

[41] “Applied Predictive Modeling | SpringerLink.” Accessed: Dec. 11, 2025. [Online]. Available: https://link.springer.com/book/10.1007/978-1-4614-6849-3

[42] A. S. Elsayad, A. I. E. Desouky, M. M. Salem, and M. Badawy, “A Deep Learning H_2 O Framework for Emergency Prediction in Biomedical Big Data,” *IEEE Access*, vol. 8, pp. 97231–97242, 2020, doi: 10.1109/ACCESS.2020.2995790.

[43] M. Sharawi, H. M. Zawbaa, E. Emary, H. M. Zawbaa, and E. Emary, “Feature selection approach based on whale optimization algorithm,” in *2017 Ninth International Conference on Advanced Computational Intelligence (ICACI)*, Doha, Qatar: IEEE, Feb. 2017, pp. 163–168. doi: 10.1109/ICACI.2017.7974502.

[44] Z. Yan, S. Wang, B. Liu, and X. Li, “Application of Whale Optimization Algorithm in Optimal Allocation of Water Resources,” *E3S Web Conf.*, vol. 53, p. 04019, 2018, doi: 10.1051/e3sconf/20185304019.

[45] S. Bag, S. K. Kumar, and M. K. Tiwari, “An efficient recommendation generation using relevant Jaccard similarity,” *Information Sciences*, vol. 483, pp. 53–64, May 2019, doi: 10.1016/j.ins.2019.01.023.

[46] G. Travieso, A. Benatti, and L. da F. Costa, “An Analytical Approach to the Jaccard Similarity Index,” 2024, *arXiv*. doi: 10.48550/ARXIV.2410.16436.

[47] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[48] T. Chen, “XGBoost: A Scalable Tree Boosting System,” *Cornell University*, 2016.

[49] S. H. Godasiaei, “A Study on Machine Learning Models’ Capability as an Alternative for CFD in Modeling Heat Transfer,” Aug. 06, 2024, *In Review*. doi: 10.21203/rs.3.rs-4690809/v1.

[50] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” *The journal of machine learning research*, vol. 13, no. 1, pp. 281–305, 2012.

[51] J. Snoek, H. Larochelle, and R. P. Adams, “Practical bayesian optimization of machine learning algorithms,” *Advances in neural information processing systems*, vol. 25, 2012.

[52] J. A. Cruz and D. S. Wishart, “Applications of Machine Learning in Cancer Prediction and Prognosis,” *Cancer Inform*, vol. 2, p. 117693510600200, Jan. 2006, doi: 10.1177/117693510600200030.

[53] G. C. Cawley and N. L. Talbot, “On over-fitting in model selection and subsequent selection bias in performance evaluation,” *The Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010.

[54] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[55] J. H. Friedman, “Greedy function approximation: A gradient boosting machine.,” *Ann. Statist.*, vol. 29, no. 5, Oct. 2001, doi: 10.1214/aos/1013203451.

[56] L. Breiman, “Bagging Predictors,” *Machine Learning*, vol. 24, no. 2, pp. 123–140, Aug. 1996, doi: 10.1023/A:1018054314350.

[57] C. E. Metz, “Basic principles of ROC analysis,” *Seminars in Nuclear Medicine*, vol. 8, no. 4, pp. 283–298, Oct. 1978, doi: 10.1016/S0001-2998(78)80014-2.

[58] *Advanced Data Mining Techniques*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. doi: 10.1007/978-3-540-76917-0.

[59] S. Visa, B. Ramsay, A. L. Ralescu, and E. Van Der Knaap, “Confusion matrix-based feature selection.,” *Maics*, vol. 710, no. 1, pp. 120–127, 2011.

[60] S. Mirjalili and A. Lewis, “The Whale Optimization Algorithm,” *Advances in Engineering Software*, vol. 95, pp. 51–67, May 2016, doi: 10.1016/j.advengsoft.2016.01.008.

[61] K. Shrestha, H. M. J. O. Rifat, U. Biswas, J.-J. Tiang, and A.-A. Nahid, “Predicting the Recurrence of Differentiated Thyroid Cancer Using Whale Optimization-Based XGBoost Algorithm,” *Diagnostics*, vol. 15, no. 13, p. 1684, July 2025, doi: 10.3390/diagnostics15131684.

[62] J. Miao and L. Niu, “A Survey on Feature Selection,” *Procedia Computer Science*, vol. 91, pp. 919–926, 2016, doi: 10.1016/j.procs.2016.07.111.

[63] W. E. Marcilio and D. M. Eler, “From explanations to feature selection: assessing SHAP values as feature selection mechanism,” in *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, Recife/Porto de Galinhas, Brazil: IEEE, Nov. 2020, pp. 340–347. doi: 10.1109/SIBGRAPI51738.2020.00053.

[64] T. Öznacar and T. Güler, “Prediction of Early Diagnosis in Ovarian Cancer Patients Using Machine Learning Approaches with Boruta and Advanced Feature Selection,” *Life*, vol. 15, no. 4, p. 594, Apr. 2025, doi: 10.3390/life15040594.

[65] Q. Mi, “Data for: USING MACHINE LEARNING TO PREDICT OVARIAN CANCER.” Mendeley, 2020. doi: 10.17632/TH7FZTBRV9.11.

[66] S. M. Ayyoubzadeh, M. Ahmadi, A. B. Yazdipour, F. Ghorbani-Bidkorpeh, and M. Ahmadi, “Prediction of ovarian cancer using artificial intelligence tools,” *Health Science Reports*, vol. 7, no. 7, p. e2203, July 2024, doi: 10.1002/hsr2.2203.

[67] Md. M. Ahamad *et al.*, “Early-Stage Detection of Ovarian Cancer Based on Clinical Data Using Machine Learning Approaches,” *JPM*, vol. 12, no. 8, p. 1211, July 2022, doi: 10.3390/jpm12081211.

[68] S. M. Ayyoubzadeh, M. Ahmadi, A. B. Yazdipour, F. Ghorbani-Bidkorpeh, and M. Ahmadi, “Prediction of ovarian cancer using artificial intelligence tools,” *Health Science Reports*, vol. 7, no. 7, p. e2203, July 2024, doi: 10.1002/hsr2.2203.

[69] G. Kaur and S. Arora, “Chaotic whale optimization algorithm,” *Journal of Computational Design and Engineering*, vol. 5, no. 3, pp. 275–284, July 2018, doi: 10.1016/j.jcde.2017.12.006.

[70] I. N. Trivedi, P. Jangir, A. Kumar, N. Jangir, and R. Totlani, “A Novel Hybrid PSO–WOA Algorithm for Global Numerical Functions Optimization,” in *Advances in Computer and Computational Sciences*, vol. 554, S. K. Bhatia, K. K. Mishra, S. Tiwari, and V. K. Singh, Eds., in Advances in Intelligent Systems and Computing, vol. 554. , Singapore: Springer Singapore, 2018, pp. 53–60. doi: 10.1007/978-981-10-3773-3_6.

[71] S. Lalwani, S. Singhal, R. Kumar, and N. Gupta, “A comprehensive survey: Applications of multi-objective particle swarm optimization (MOPSO) algorithm,” *Transactions on combinatorics*, vol. 2, no. 1, pp. 39–101, 2013.