# AN ML-BASED DECISION SUPPORT SYSTEM FOR RELIABLE DIAGNOSIS OF OVARIAN CANCER BY LEVERAGING EXPLAINABLE AI

*Report submitted to the SASTRA Deemed to be University as the requirement for the course*

**CSE599 – Project Phase – I**

*Submitted by*

**THANISH S**
**(Reg. No.: 125162022, MTech – Artificial Intelligence and Data Science)**

**November 2024**

**SCHOOL Of COMPUTING**

**THANJAVUR, TAMIL NADU, INDIA – 613 401**

SASTRA

ENGINEERING · MANAGEMENT · LAW · SCIENCES · HUMANITIES · EDUCATION

DEEMED TO BE UNIVERSITY

(U/S 3 of the UGC Act, 1956)

THINK MERIT | THINK TRANSPARENCY | THINK SASTRA

THANJAVUR | KUMBAKONAM | CHENNAI

**SCHOOL OF COMPUTING**

**THANJAVUR – 613 401**

## Bonafide Certificate

This is to certify that the report titled "An ML-based decision support system for reliable diagnosis of ovarian cancer by leveraging explainable AI" submitted as a requirement for the course, CSE599 Project Phase - I for MTech. is a Bonafide record of the work done by Mr. THANISH S (Reg. No.125162022, Artificial Intelligence and Data Science) during the academic year 2024-25, in the School of Computing, under my supervision**.**

**Signature of Project Supervisor:**

**Name with Affiliation:** **Dr. PLK Priyadarsini, Associate Professor, SOC, SASTRA**

**Date:** **11/11/24**

Mini Project *Viva voce* held on _____

**Examiner 1** **Examiner 2**

# Acknowledgements

I would like to thank our Honorable Chancellor **Prof. R. Sethuraman** for providing us with an opportunity and the necessary infrastructure for carrying out this project as a part of our curriculum.

I would like to thank our Honorable Vice-Chancellor **Dr. S. Vaidhyasubramaniam** and **Dr. S. Swaminathan,** Dean**,** Planning & Development, for the encouragement and strategic support at every step of our college life.

I extend our sincere thanks to **Dr. R. Chandramouli,** Registrar, SASTRA Deemed to be University for providing the opportunity to pursue this project.

I extend our heartfelt thanks to **Dr. V. S. Shankar Sriram**, Dean, School of Computing, **Dr. R. Muthaiah**, Associate Dean, Research, **Dr. K.Ramkumar**, Associate Dean, Academics, **Dr. D. Manivannan**, Associate Dean, Infrastructure, **Dr. R. Algeswaran**, Associate Dean, Student Welfare.

My guide **Dr. P.L.K Priyadarsini**, Associate Professor, School of Computing was the driving force behind this whole idea from the start. His deep insight in the field and invaluable suggestions helped me in making progress throughout our project work. I also thank the project review panel members for their valuable comments and insights which made this project better.

I would like to extend our gratitude to all the teaching and non-teaching faculties of the School of Computing who have either directly or indirectly helped me in the completion of the project.

I gratefully acknowledge all the contributions and encouragement from my family and friends resulting in the successful completion of this project. I thank you all for providing me an opportunity to showcase my skills through project.

# List of Figures



# List of Tables

# Abbreviations

| | |
|---|---|
| AI | Artificial Intelligence |
| MICE | Multiple imputation using chained equations |
| IQR | Interquartile Range |
| SHAP | SHapley Additive exPlanations |
| SMOTE | Synthetic Minority Oversampling Technique |
| ROMA | Risk of Ovarian Malignancy Algorithm |
| OC | Ovarian Cancer |
| RMI | Risk of Malignancy Index |
| DNA | Deoxyribonucleic acid |
| XAI | Explainable AI |

## ABSTRACT

Ovarian cancer is amongst the deadliest gynaecological cancers, mainly due to the fact that it is initiated asymptomatically in early stages and diagnosed much later, hence extremely high mortality rates. There is a crucial demand for the early detection of the disease with good accuracy, and therefore, the current project hopes to fill in this gap by building onto the existing systems of diagnostics that use clinical and histopathological data analysis. Some traditional ML algorithms were chosen with XGBoost being the ultimate choice of model due to their high performance, such as Random Forest, SVM, and Decision Trees. Model performance metrics - accuracy, precision, recall, and F1-score - are put in place to ensure reliable diagnostics, and Explainable AI techniques like SHapley Additive exPlanations (SHAP) provide understanding of model decisions. Several innovations have been introduced into this project that take care of limitations in existing systems. Advanced feature selection techniques, including Genetic Algorithms, are then used to identify the most important key biomarkers, refine model inputs, and be able to predict much more accurately. Other ensemble approaches, like bagging, boosting, and stacking, are applied for enhancing robustness and accuracy. Hyperparameters are tuned through GridSearch, and XGBoost has been optimized for best possible diagnostic performance of the model. A full web application has been developed to further ease the process of interaction by users. This application integrates a chatbot developed using the Mistral 7B model, which helps to assist users by asking for biomarkers, predicting the stage of cancer, and further offering suggestions for hospitals, diet, precautions, and treatments. In this manner, the project finally produces a diagnostic tool robust, understandable by all, which is aimed at making early detection possible so that proper treatment can be evolved for ovarian cancer patients.

# CHAPTER 1

## 1. INTRODUCTION

Ovarian cancer (OC) is a gynaecological cancer that involves rapid cell mutation and abnormal tissue growth; metastatic spread can often occur if it's not diagnosed early. It ranks the second leading gynaecologic cause of death following its reason for low survival rates resulting from delayed diagnoses and lack of proper tools in screening. Statistics in 2018 report over 295,000 new cases and approximately 185,000 deaths worldwide due to OC. Early stages are crucial because the five-year survival rate tapers precipitously with increasing stages: it is close to 92.6% for stage I but drops to about 30.3% for stage IV. In traditional OC diagnostics, various markers like CA125, CA72-4, and HE4 are used along with imaging studies. Still, these methods often involve less sensitivity and specificity that makes it difficult to make a reliable early detection.

Machine learning, one of the most innovative areas developed within disease diagnosis. The area that possesses significant potential for enhancing accuracy in diagnosis will be realized through sophisticated pattern recognition and predictive modelling. The proposed project represents an ML-based framework for enhancing the accuracy of detection for OC. Focus was given to the use of XGBoost as the principal classifier because it is strong, accurate, and very powerfully generalized. Using an incredibly large dataset, our work trains and tests the model on 171 malignant and 178 benign cases. It further extends by incorporating Shapley Additive exPlanations (SHAP) for interpreting the model predictions so as to fill the system with more clarity and trust for clinical applications. By applying GA and SHAP together in the current work, we were able to identify and differentiate crucial biomarkers for the premenopausal and postmenopausal populations, leading to an even more personalized diagnostic approach. Using GridSearch to tune the hyperparameters, it obtained accuracy rates of 89.6% and 95.8% for pre and postmenopausal cases, respectively.

In addition, the project includes a friendly web application, as well as a chatbot powered by the Mistral 7B chatbot for interaction with the patient. The chatbot will collect information on biomarkers, provide medical advice, and make recommendations about healthcare facilities, treatments, and dietary recommendations. The hybrid framework attempts to bring together some of the main leaders in terms of predictive accuracy and model interpretability into a trustworthy, transparent tool for the diagnosis of OC.

## 1.1 Summary of the Base paper:

Based on classical models, for instance, the Risk of Ovarian Malignancy Algorithm (ROMA) and the Risk of Malignancy Index (RMI), existing approaches are reviewed for the diagnosis of OC. ROMA, through HE4 and CA125 biomarkers, stratifies the risk of OC; with a sensitivity of 94.3%, this makes it one of the most popular methods to determine the risk of OC. RMI integrates ultrasound scores with CA 125 levels and menopausal status to predict malignancy; it achieves a sensitivity of 84.6%. Overall, the studies have indicated that ROMA has an edge in terms of accuracy and objectivity as well. Research studies on individual biomarkers, mainly HE4, has also suggested the same view; HE4 is said to be highly sensitive for differentiating ovarian masses and has specific relevance to premenopausal women. These multi-marker approaches, especially the combined usage of HE4, CA125, progesterone (Prog), and estradiol (E2), have been particularly promising when used with ML algorithms for feature selection as reported earlier, where HE4 and CEA were found to be critical predictors with an accuracy of 89.0% for OC.

Even though there are significant advances in these models, the major limit in models so far developed is that these are difficult to interpret and hence not clinically applicable. The base paper emphasizes the gap that is recorded in the application of XAI framework diagnosis because the models are expected to offer both accurate predicament and the transparency in reasoning. Thus, this research study fills this gap through developing an XAI-enhanced ML framework, making use of GA for feature selection as well as SHAP for interpretability, thus intending to diagnose and to bring about vital transparency-based needs essential in clinical settings.

## 1.2. Merits of the Base Paper:

The base paper demonstrates that the ROMA model possesses a strong predictive capability, with a sensitivity of 94.3%, very apt for risk stratification.

In-depth Analysis of Biomarkers: It contains both single-type and multi-type biomarkers, hence providing in-depth analysis of various combinations of biomarkers.

Machine Learning Innovation: The study manifests the application of feature selection techniques wherein Minimum Redundancy Maximum Relevance technique has been used to obtain tremendous enhancements in accuracy compared to classical approaches.

Focus on Explainable AI: The base paper identifies the necessity of model transparency and lays the foundation for the addition of explainable AI within the framework of OC diagnosis, which is required for clinical application.

Segregation Based on Menopausal Status: The paper throws light on the significance of biomarkers based on menopausal status, which leads to more personalized diagnosis approaches.

## 1.3. Demerits of Base Paper

A limited interpretability framework: It doesn't discuss or develop an explainable AI framework in detail, hence keeps it within the realm of further scope for improvement.

Data limitations: Most of the studies quoted here rely on constrained datasets; sometimes, these may potentially limit the generalizability of findings for other types of populations.

Lack of Consistency in the Relevance of Biomarkers: The article reflects some inconsistency in the results in relation to the effectiveness of the biomarkers, such as different sensitivities by HE4 and CA125 across different studies, which leads to inconsistent feature selection.

The models presented do not provide an interactive system where patients and physicians can communicate directly right away, so their practical use is minimal.

Limited Emphasis on Hyperparameter Optimization: The base paper has not given consideration to any techniques to tune hyperparameters for optimal ML model performance. Such factors would optimize the accuracy of diagnosis.

# CHAPTER 2

## 2. RELATED WORKS

Among the early screening tools developed to facilitate an easier diagnosis of ovarian cancer (OC) through the use of tumour biomarkers, one of the most profound approaches is the Risk of Ovarian Malignancy Algorithm (ROMA), proven to have predictive accuracy, particularly if HE4 and CA125 biomarkers are used in classification to stratify patients to either a high- or low-risk group for OC. The ROMA algorithm is good at stratifying patients and has, therefore, become the go-to evaluation of the risk of OC.

Another well-established is the Risk of Malignancy Index (RMI), which takes a combination of ultrasound scores with CA125 serum levels and menopausal status to provide a dependable preoperative diagnosis of malignancy in women. Though RMI is effective, ROMA's biomarker-driven approach is simpler and more objective. A study compared the two, showing that in diagnosing OC, ROMA demonstrated a greater sensitivity at 94.3% than RMI with 84.6%.

Biomarkers have been specifically targeted by other studies like HE4 and CA125. A study on 128 women with HE4 was highly sensitive in the differentiation of ovarian masses. In another study, HE4 was found to be the most accurate predictor of OC wherein ROC-AUC values were registered at 0.9 for HE4, 0.893 for ROMA, 0.874 for USG, and 0.794 for CA125. A meta-analysis reported that HE4 particularly served in diagnosing OC in premenopausal women whereas, ROMA and CA125 showed significant improvement in the postmenopausal diagnosis. Interestingly, another study presented different outcomes. It obtained elevated CA125 sensitivity for the detection of premenopausal OC and established new cut-off ROMA values (14.9% in case of premenopausal and 33.4% for postmenopausal) in order to enhance their prediction abilities.

The last developments in the OC diagnostic models are multi-marker approaches. One model has included HE4, CA125, progesterone (Prog), and estradiol (E2) testing that emphasizes the possible role of steroid hormones for OC detection. ML also seems to be a useful tool as one of the first studies in developing an ML-based prediction system for OC. By using the feature selection Minimum Redundancy Maximum Relevance, they reduced the feature set to 10 and found HE4 and CEA to be critical predictors, thereby achieving accuracy of 84.7%, which surpassed ROMA at 79.6%. Later work on multiple ML algorithms indicated that the classifier

LGBM could obtain the highest accuracy of 88%. Few studies have focused on the use of explainable AI, or XAI, in interpreting model predictions. Most prior work lacks an interpretation framework that is a limitation of applying such models in clinical settings where understanding model reasoning is crucial. There is also conflicting information regarding which are the more relevant biomarkers, which makes it difficult to decide on a more consistent approach for selection of biomarkers in OC diagnosis. This study will bridge this gap by creating an XAI framework that utilises a GA for feature selection and SHAP for explanations along with applying classification using XGBoost. Our proposed approach will not only make available a sound ML-based diagnostic tool but also allow transparency in model predictions. Thus, it will help clinicians take well-informed and reliable decisions.

# CHAPTER 3

# 3. DATA PREPROCESSING AND MODELLING

### 3.1. Data collection

The source of data for this research was the Third Affiliated Hospital of Soochow University. It was a collection of patients' histories for 349 cases with ovarian cancer (OC) who were operated on through surgical resection from 2011 to 2018. Of these patients, 171 were malignant and the other 178 were benign. Significantly, no patient involved in this study underwent any radiotherapy or chemotherapy before surgery, hence there was least probability of confounding treatment effects.

This dataset contains 49 feature variables, which are categorized into the following: blood routine tests, tumour biomarkers, general chemistry tests, and demographic information of the patients. The target variable is binary in the sense that this variable informs whether the tumor is malignant or benign. Differences between the malignant and benign groups were tested using the Mann-Whitney U Test, applied to each attribute. Results, including p-values along with summary statistics such as mean, median, and range for each attribute.

| Feature Name | Acronym | Unit | No. of instances | Mean | Median | Range | p-value |
|---|---|---|---|---|---|---|---|
| Demographics | | | | | | | |
| Age | | | 349 | 45.05 | 45 | 15-83 | 0.001 |
| Benign Ovarian Tumors | | | 178 | | | | |
| Ovarian Cancer | | | 171 | | | | |
| Premenopausal | | | 230 | | | | |
| Postmenopausal | | | 119 | | | | |
| Blood Routine Test | | | | | | | |
| Basophil Cell Count | BASO# | 10^9/L | 349 | 0.03 | 0.03 | 0.00-0.12 | 0.312 |
| Basophil Cell Ratio | BASO% | % | 349 | 0.48 | 0.4 | 0.00-1.94 | 0.054 |
| Eosinophil Count | EO# | 10^9/L | 349 | 0.07 | 0.05 | 0.00-0.40 | 0.421 |
| Eosinophil Ratio | EO% | % | 349 | 1.12 | 0.8 | 0.00-7.60 | 0.134 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Haematocrit | HCT | L/L | 349 | 0.38 | 0.39 | 0.22-0.57 | 0.033 |
| Haemoglobin | HGB | g/L | 349 | 125.34 | 127 | 61.80-189.00 | 0.001 |
| Lymphocyte Count | LYM# | 10^9/L | 349 | 1.56 | 1.5 | 0.35-3.49 | 0.001 |
| Lymphocyte Ratio | LYM% | % | 349 | 26.07 | 26.6 | 3.90-51.60 | 0.001 |
| Mean Corpuscular Haemoglobin | MCH | Pg | 349 | 28.78 | 29.3 | 17.70-36.80 | 0.001 |
| Mean Corpuscular Volume | MCV | fL | 349 | 88.07 | 89 | 61.00-107.90 | 0.414 |
| Mononuclear Cell Count | MONO# | $10^9$/L | 349 | 0.36 | 0.32 | 0.07-0.97 | 0.001 |
| Monocyte Ratio | MONO% | % | 349 | 5.58 | 5.43 | 0.30-21.30 | 0.068 |
| Mean Platelet Volume | MPV | fL | 347 | 10.04 | 10.3 | 5.06-14.50 | 0.407 |
| Neutrophil Ratio | NEU | % | 258 | 66.58 | 66.75 | 37.20-92.00 | 0.001 |
| Thrombocytocrit | PCT | L/L | 347 | 0.25 | 0.24 | 0.07-0.69 | 0.001 |
| Platelet Distribution Width | PDW | % | 347 | 14.33 | 13.7 | 8.80-22.80 | 0.002 |
| Platelet Count | PLT | 10^9/L | 349 | 255.43 | 236 | 74.00-868.00 | 0.001 |
| Red Blood Cell Count | RBC | 10^12/L | 349 | 4.36 | 4.37 | 2.62-6.74 | 0.031 |
| Red Blood Cell Distribution Width | RDW | % | 349 | 13.55 | 13.1 | 10.92-22.20 | 0.376 |
| General Chemistry | | | | | | | |
| Anion Gap | AG | mmol/L | 348 | 19.32 | 19.85 | 6.20-33.33 | 0.339 |
| Albumin | ALB | g/L | 339 | 41.08 | 42 | 22.00-51.50 | 0.001 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Alkaline Phosphatase | ALP | U/L | 339 | 77.09 | 71 | 26.00-763.00 | 0.001 |
| Alanine Aminotransferase | ALT | U/L | 339 | 18.01 | 15 | 4.00-86.00 | 0.353 |
| Aspartate Aminotransferase | AST | U/L | 339 | 19.11 | 17 | 7.00-78.00 | 0.001 |
| Blood Urea Nitrogen | BUN | mmol/L | 349 | 4.01 | 3.83 | 1.12-10.19 | 0.449 |
| Calcium | Ca | mmol/L | 349 | 2.39 | 2.47 | 0.92-2.83 | 0.001 |
| Chlorine | CL | mmol/L | 349 | 100.81 | 100.9 | 84.60-109.40 | 0.34 |
| Carbon Dioxide-combining Power | CO2CP | mmol/L | 348 | 24.28 | 24.05 | 16.20-34.30 | 0.329 |
| Creatinine | CREA | µmol/L | 349 | 64.25 | 63.3 | 38.20-114.00 | 0.06 |
| Direct Bilirubin | DBIL | µmol/L | 339 | 3.13 | 2.8 | 0.90-12.10 | 0.001 |
| Gama Glutamyl transferase | GGT | U/L | 339 | 21.31 | 16 | 4.00-176.00 | 0.007 |
| Globulin | GLO | g/L | 339 | 30.18 | 30.1 | 14.10-47.60 | 0.001 |
| Glucose | GLU. | mmol/L | 349 | 5.33 | 5.08 | 3.57-12.44 | 0.014 |
| Indirect Bilirubin | IBIL | µmol/L | 339 | 5.96 | 5.4 | 1.00-28.40 | 0.001 |
| Kalium | K | mmol/L | 349 | 4.39 | 4.37 | 3.08-5.40 | 0.908 |
| Magnesium | Mg | mmol/L | 349 | 0.98 | 0.97 | 0.65-1.37 | 0.571 |
| Natrium | Na | mmol/L | 349 | 140.49 | 140.5 | 125.1-150.7 | 0.001 |
| Phosphorus | PHOS | mmol/L | 349 | 1.12 | 1.12 | 0.57-1.75 | 0.641 |
| Total Bilirubin | TBIL | µmol/L | 339 | 9.09 | 8.4 | 2.5-38.3 | 0.001 |
| Total Protein | TP | g/L | 339 | 71.08 | 72.5 | 32.9-86.8 | 0.015 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Uric Acid | UA | µmol/L | 349 | 243.71 | 235.4 | 96.0-632.0 | 0.982 |
| Tumour Marker | | | | | | | |
| Alpha - fetoprotein | AFP | ng/mL | 327 | 11.82 | 2.28 | 0.61-1210.0 | 0.012 |
| Carbohydrate Antigen 125 | CA125 | U/mL | 332 | 350.38 | 44.68 | 3.75-5000.0 | 0.001 |
| Carbohydrate Antigen 19-9 | CA19-9 | U/mL | 325 | 46.73 | 14.2 | 0.6-1000.0 | 0.2 |
| Carbohydrate Antigen 72-4 | CA72-4 | U/mL | 109 | 10.17 | 2.37 | 0.2-158.5 | 0.001 |
| Carcinoembryonic Antigen | CEA | ng/mL | 327 | 3.31 | 1.33 | 0.2-138.8 | 0.103 |
| Human Epididymis Protein 4 | HE4 | pmol/L | 329 | 183.95 | 53.27 | 16.71-3537.6 | 0.001 |

Table 1. Statistical summary of the data

### 3.2. Data preprocessing

Raw dataset underwent pre-processing to remove anomalies so that missing values and outliers would be taken care of to ensure good quality data for modelling

Data cleaning: The column of ID has been excluded since it doesn't add any predictive value, and extra characters on data records were deleted for data consistency.

Missing Data Handling: As depicted in Table 1, the following feature variables had some missing entries, namely NEU, MPV, and CA19-9. Since the size of the dataset is not large, it's not feasible to drop any variables or samples that contain missing values as it has the possibility to shrink the size of the dataset further. Hence, an MICE algorithm called Multivariate Imputation by Chained Equations is used for imputation of missing values. MICE is a robust imputation algorithm, as it uses regression modelling to predict and replace missing entries based on considering each attribute with missing data as a function of other variables.

Outliers Handling: There were some outliers in particular attributes, which are extreme data points that differ significantly from other observations. Outliers may occur either due to

mistakes in measurement or due to natural variations in the population being considered. In any case, they induce bias and lower the interpretability of the model. Fig. 1 depict the box plots that show outliers for each of the features. The outlier values were scaled back within the limits defined as ±1.5 times the interquartile range (IQR) of the quartiles by truncation.

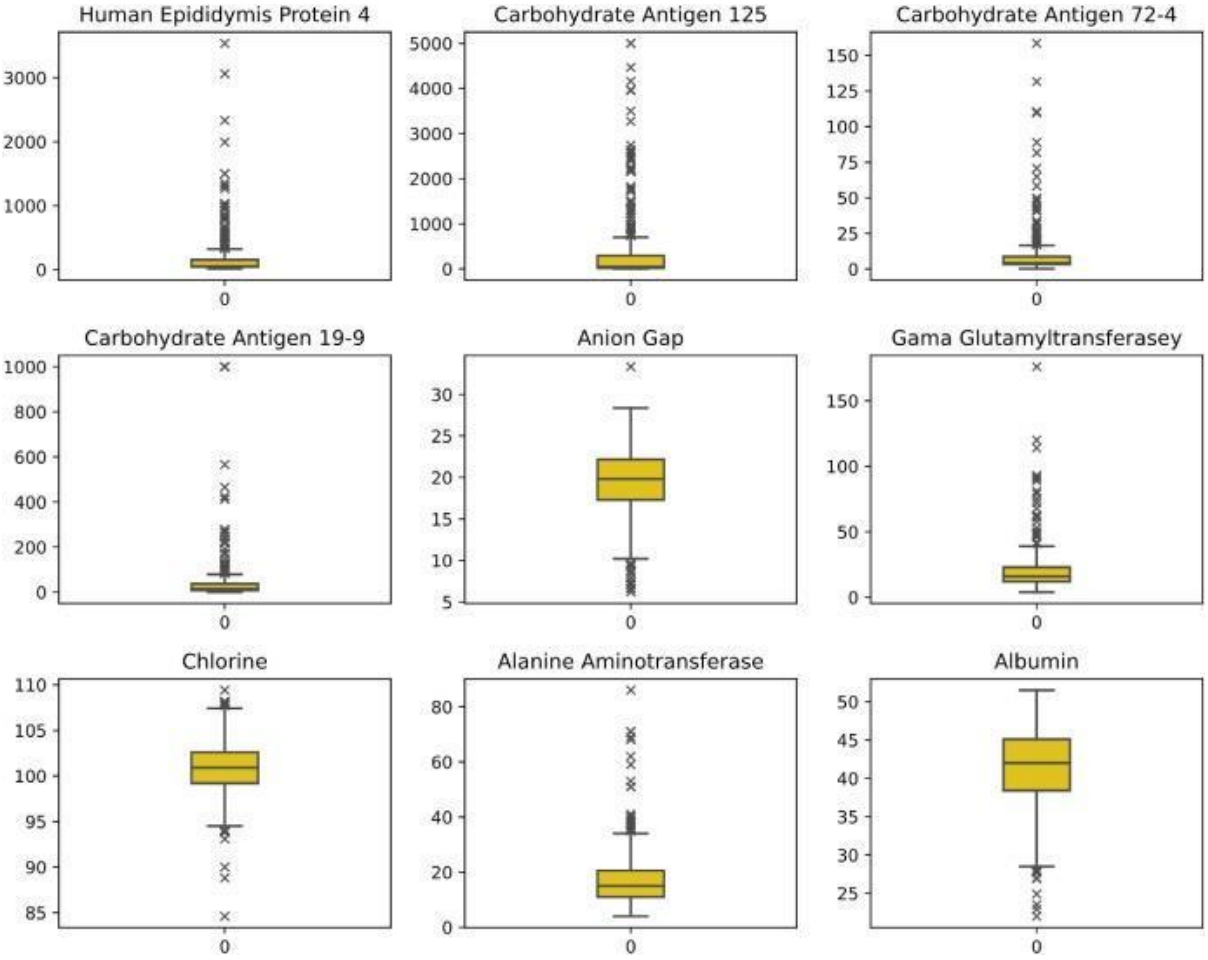



Fig. 1. Box plot representing the presence of outliers in different features.

Normalization: All the feature variables were scaled to a common range by centring all the variables through z-score transformation.

$$z = \frac{x - \mu}{\sigma}$$

where, $\mu$ = mean, and $\sigma$ = standard deviation.

### 3.3 Proposed Methodology

The processed dataset was applied to the ML pipeline. The FS technique was used to identify the most relevant features. A GA was selected based on its efficiency in searching for the optimal feature subset. Since it is a wrapper method, a GA evaluates the performance of a classification model to make iterative selection of the best feature combination. The classification was carried out using the XGBoost classifier due to its strength and performance in dealing with structured data.

Given the importance of the status of menopause as an important risk factor for ovarian cancer (OC), separate models were developed for premenopausal and postmenopausal women. Different cut-off values of biomarkers are often used for these two groups. Therefore, the feature selection process was conducted separately for both populations to ensure the selection of the most relevant features for each group.

A 10-fold cross-validation strategy was utilized to evaluate the performance of the XGBoost model. It would train and test the model on several subsets of the data to make a better predictor of the overall performance of the model. Seven metrics were utilized to check how well the classifier is effective.

SHAP, or SHapley Additive exPlanations, is used to explain how features contribute to the model predictions to increase interpretability. Different visualization techniques were used both for global interpretability and local interpretability. Partial dependence plots and variable importance plots for achieving global interpretability, which are measures of overall significance in terms of feature contribution. In creating a deeper understanding regarding the decision-making of the model over certain patients, individual predictions were analyzed with the help of a SHAP force plot. SHAP has been applied further to look into the wrong predictions made by the model, which shed more light on the areas that need improvement in the model.

Finally, performance comparison was done with the algorithm ROMA and other previously published studies, ensuring that the new approach improves the method both in accuracy of prediction and interpretability. A schematic of the proposed methodology is shown in Fig. 2.

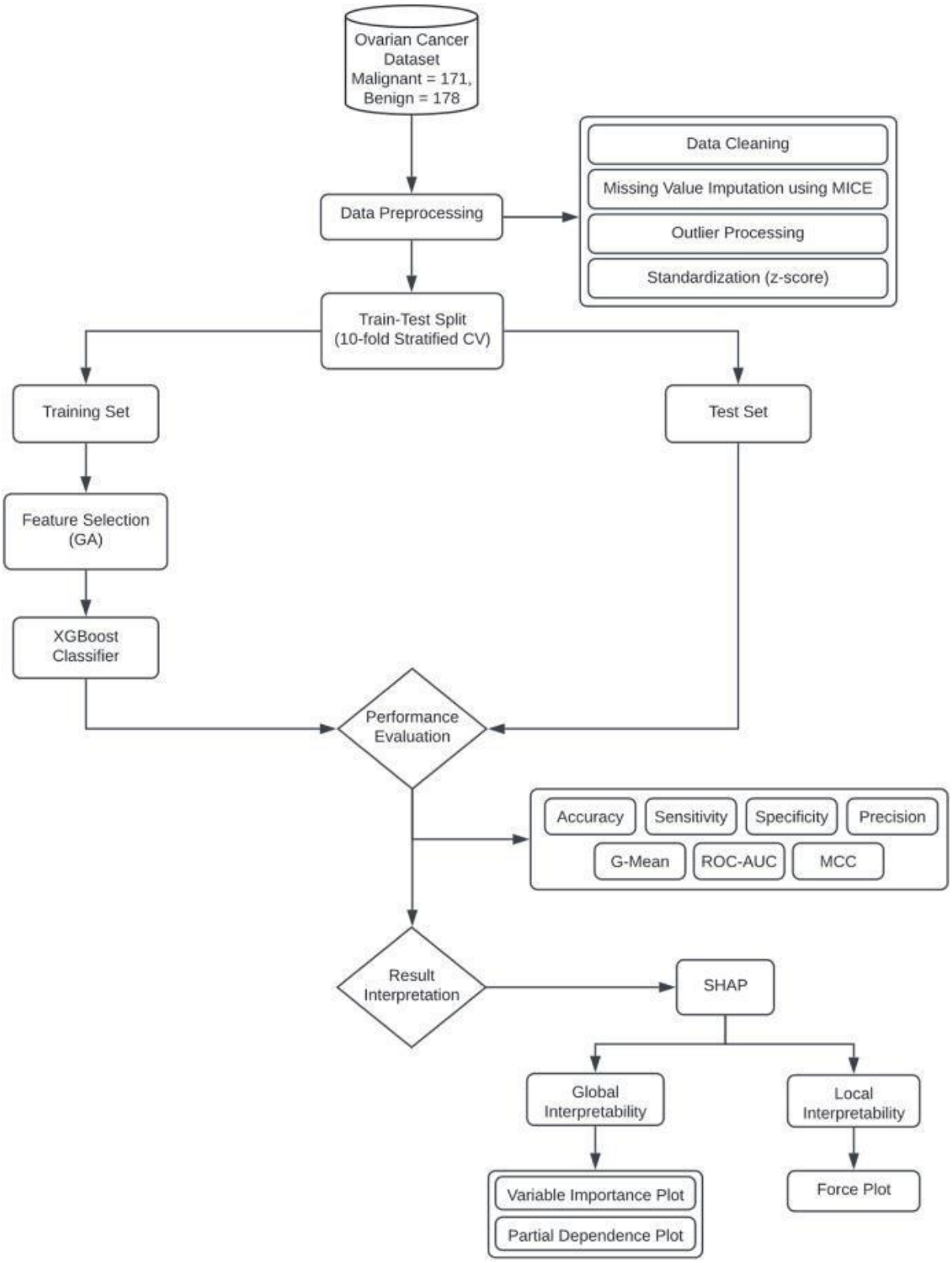



Fig 2: Outline of the proposed framework.

### 3.4. Feature selection using genetic algorithm

The use of feature selection techniques is, therefore important since it allows for the selection of the most relevant features, neglecting redundant or irrelevant ones, and thus improves machine learning performance and efficiency. Among the most developed and widely used optimization techniques used for feature selection, there is the genetic algorithm. The biological principles that inspire this paper-biological evolution-are leveraged in a technique called genetic algorithms using operators like crossover, mutation, and selection to evolve the population of potential candidate solutions as well as select the most relevant feature(s) from the dataset.

Genetic Algorithms Process: Given a population of candidate solutions, each a subset of features, new offspring (subsets of features) are generated through such processes as recombination (crossover) and mutation. It is a solution in the process of evolution. The initial population consists of randomly chosen subsets of features. At every iteration, a new generation of feature subsets is produced. All individuals in the population are examined using an objective function that assigns the fitness value based on how well the given feature set performs.

It is the tournament selection mechanism which is the backbone of GA, where a more fit individual is selected based on a stochastic mechanism with a higher probability of producing offspring. For that reason, for cross-over, the XOR operator applies the XOR operation between the parent chromosomes to produce offspring which are different from their parents. It makes more random mutations, as it randomly tweaks the genes of the chromosomes; hence, this makes the search much exploratory on the solution space. The elitist selection mechanism ensures that the best individuals without any modification pass on to the next generation. It continues until the GA converges to an optimal or near-optimal subset of features.

Table 2 indicates the parameter settings applied during the design of GA in our research. A grid search method was conducted to identify the optimal value setting for the parameters. The grid search is shown below:

-Population size: This defines the number of individual solutions that are found in every generation. Large population sizes improve exploration at the expense of large increases in computational cost. Thus, there is a trade-off between exploration and the associated increase in computational cost.

- Probability of crossover: The chance that two parents swap their genetic materials at reproduction. It is the parameter that controls the balance between exploration and exploitation. The higher the probability of crossover, the more inclined it is towards convergence whereas a low value of the probability encourages exploration by creating offspring with higher variability.

- Mutation probability: It is the probability that random changes will occur in the genetic material of an individual. The mutation probability ensures that the diversity of the population remains high and the algorithm does not reach local optima. It helps explore other regions within the solution space.

- Independent probability for each attribute to be crossed over: It specifies the probability of every single attribute to undergo a crossover operation independent of the changes attributed to the other attributes. This makes the crossover uniform and balanced in all the attributes, thus offering an excellent exploration of attribute combination.

- Independence of every attribute: The individual probability of each attribute having a mutation: The probability that each attribute, on its own, is randomly mutated. These are meant to help encourage diversity and to efficiently explore the permutations of the different attributes.

Tournament size: This is the number of people selected from the population to compete in a tournament for the choice of individual to reproduce. The study simulated three such picks that are done in a random manner, and the fittest picked gets to win. This presents a stable middle route in selection pressure to choose between those pickups with high performances yet maintain the variation in the population.

This allows the configuration, in combination with GA's iterative optimization, to identify those features most relevant in producing accurate predictions in the context of ovarian cancer detection.

| Parameters | Value |
|---|---|
| Number of populations | 100 |
| Probability of crossover | 0.5 |
| Probability of mutation | 0.2 |
| Independent probability for each attribute to be exchanged | 0.5 |
| Independent probability for each attribute to be mutated | 0.5 |
| Tournament size | 3 |
| Number of generations | 20 |
| Objective function | Accuracy score |

Table 2. GA parameters.

### 3.5. ML algorithm - XGBoost classifier

Ensemble learning is an approach in which multiple classification algorithms, known as base learners, use the idea of creating a more powerful predictive model as a whole than one can produce individually. Even though individual base learners, often termed weak learners, are prone to high noise and suffer from overfitting, a well-crafted ensemble of them can create a more robust model. Among the ensemble techniques available, the two popular ones are bagging and boosting. In bagging, several weak learners are trained to different subsets of the data and their predictions are aggregated through voting or averaging to give the final prediction. Boosting primarily focuses on training a sequence of learners where each learner attempts to correct the errors of its predecessors.

Boosting algorithms are highly popular for improving the accuracy of prediction, in which most of the boosters are better than other traditional classifiers. However, out of so many boosting algorithms, AdaBoost and Gradient Boost are the most prominent ones.

XGBoost is an advanced derivative of the Gradient Boost algorithm. Its main popularity could be due to the high performance achieved in carrying out a large range of machine learning activities. XGBoost also works as a parallel learner and can handle vast amounts of data. The biggest advantage of XGBoost is that it can very well manage the complexity of models and avoid overfitting due to regularization techniques that help to get correct predictions and improve upon performance. Moreover, it's very highly optimized for computation and memory usage, which accelerates faster acceleration times than other gradient algorithms in training times.

These properties make XGBoost highly applicable in the domain of machine learning due to its fast processing, good performance, and flexibility. Here, we have used `XGBClassifier` from Python's "xgboost" package with all default parameters.

### 3.6. Hyperparameter Tuning and Model Evaluation Report

In this study, after hyperparameter tuning, the XGBoost classifier is used to optimize the performance of the target variable. These procedures include SMOTE technique-based balancing of the dataset, splitting into training and testing sets, and standardization of features. The aim is to solve overfitting problems and class imbalance issues for better accuracy.

#### 3.6.1. SMOTE for Class Imbalance

It uses SMOTE which balances the class; this is generally the problem in many real-world datasets where representatives of one class are very few. It creates synthetic samples using SMOTE so that it may balance out the proportion of the minority class so that the classifier learns equally from all classes for better prediction.

Data Partitioning and Normalization: The set was divided into the training and testing sets in a stratified manner, so both classes were represented proportionately in both subsets. The training set was used for training the model, while the testing set was kept aside for evaluation. Feature standardization was also done so that all features could be scaled to have zero mean and unit variance. Standardization is indispensable for algorithms like XGBoost because it aids in faster convergence and more reliable performance.

#### 3.6.2. Grid Search Hyperparameter Optimization

The XGBoost model was tuned through doing a grid search of several hyperparameters. In doing so, one had to consider the number of estimators or trees, the learning rate, the maximum depth of trees, minimum child weight, and the subsample ratio. Such a search was cross-validated and, due to the given model, was tried to validate the combination of good hyperparameters with accuracy as the criterion.

#### 3.6.3 Model Training with Early Stopping

Once the best hyperparameters are identified, the model is trained with early stopping. Early stopping stops training if the validation score does not improve for a specified number of iterations. This makes it impossible to overfit because the model does not continue training

when its performance is optimal. The best hyperparameters were used to train the final model, and the test set was evaluated.

### 3.6.4. Model Evaluation

Accuracy and a classification report along with ROC-AUC score were estimated after training the model in order to test the proficiency of the trained model. A general metric to check how well the model performed is accuracy. ROC-AUC score provides an insight into the model's ability to distinguish classes. The classification report comprises detailed performance metrics such as precision, recall, and F1-score that give a comprehensive look at how well the model handles both classes.

This model achieved very high accuracy and displayed an impressive ROC-AUC score, as the hyperparameter tuning with application of SMOTE proved influential in enhancing the performance of the model. The classification report also assured that the model was successful in identifying positive as well as negative cases at appropriate balance between precision and recall.

Appropriate application of the process of hyperparameter tuning, combined with the usage of SMOTE in balancing the dataset, has resulted in highly improved performance by the XGBoost classifier. These complementary techniques have enabled the model to handle class imbalance, avoid overfitting, and make true predictions. The following evaluation metrics showed that the model is capable of generalizing well to unseen data; hence the model will be suitable for application in real-world applications.

## 3.7. Explainable AI with SHAP

A highly advanced ML algorithm might create very accurate predictions, but the black-box nature of such predictions severely limits their applicability to real-world applications. It is particularly disturbing in high-stake fields like the medical industries where the prediction could be a source of critical devastation if the machine goes wrong. Thus, the direction has been more towards explaining ML models in ways that can enhance their trustworthiness and accountability. Explainable AI (XAI) constitutes an emerging trend that is dedicated to developing the processes and methodologies meant to make the meanings of what the ML algorithms produce intelligible. It is this intelligibility, thus, which may render AI systems much more reliable and effective as clinical decision-support tools.

One of the instruments in XAI is SHAP, based on a simple idea from game theory. SHAP calculates Shapley values for each sample and feature in the dataset; however, this calculation is model-agnostic. The Shapley value was actually designed within cooperative game theory as a way to distribute both costs and benefits between cooperative players. SHAP leverages these values to provide both global and local interpretations. The global interpretations explain the feature relationships and the target variable for the entire data, while the local interpretations explain how the individual contribution of features contributes to a prediction that has been made for an instance. The SHAP importance of features can vary between the global and local contexts, which makes SHAP flexible as well as a very insightful tool.

Using SHAP, one can create personalized risk prediction models, thus aiding doctors significantly in more accurate and interpretable predictions. Such local explanations help understand why a particular patient receives a specific prediction and how the various attributes contribute to the final outcome.

The SHAP Python module has several techniques for the interpretation of model predictions. These include partial dependence plots, that show the marginal effect of one or two features on the predicted outcome, variable dependence plots, that display the relationship between predictor variables and target variables; force plots, which help to interpret individual predictions for different patients. Visualization techniques are important in terms of explaining how the model ends up predicting its results and are significant in the context of assuring the transparency as well as the reliability of ML models when applied in high-risk scenarios.

### 3.8. ROMA score

ROMA is a popular approach used in the classification of pelvic masses. It is calculated using the concentrations of the serum biomarkers HE4 and CA125 in conjunction with the menopausal status of the patient. The predictive index of ROMA is determined using the following formulae:

Pre-menopausal Predictive Index (PI) = $-12.0 + 2.38 \times \ln(\text{HE4}) + 0.0626 \times \ln(\text{CA 125})$

Post-menopausal Predictive Index (PI) = $-8.09 + 1.04 \times \ln(\text{HE4}) + 0.732 \times \ln(\text{CA 125})$

Predicted probability of the ROMA score = $\frac{e^{\pi}}{1+e^{\pi}} * 100$

The ROMA score provides a threshold for determining the risk of ovarian cancer based on menopausal status:

- For premenopausal women, a ROMA score of 13.1% is the cut-off value indicating a high risk of ovarian cancer.
- For postmenopausal women, the high-risk threshold is 27.7%.

These thresholds help clinicians differentiate between benign and malignant pelvic masses. During the experimentation phase in this study, the ROMA scores for each patient in the dataset were calculated using the appropriate formulas, and the results were compared with those from the proposed approach to assess the accuracy and effectiveness of the new model in predicting ovarian cancer.

The integration of a chatbot in the detection system of ovarian cancer is to enhance user experience and offer personalized support for the patients and healthcare professionals at every stage of the diagnostic process. The intelligent assistant can collect real-time data, determine the related risk, and make recommendations based on patient-specific information, thereby making the model easier to understand as well as more accessible.

### 3.9. Overview of Chatbot Functionality

The main task of the chatbot is to collect key values of biomarkers and menopause status of patients. Taking the above inputs, it computes the ROMA score to find the possibility of ovarian cancer. The characteristics that the chatbot offers are as follows:

#### 3.9.1. Data Collection:

The patients are required to input the essential values of biomarkers such as HE4 and CA125, along with the menopausal status whether she is pre-menopausal or post-menopausal. The data is used for calculating the ROMA score.

#### 3.9.2. Real-Time Risk Computation:

Once the input is received, the chatbot will make real-time calculations to determine the ROMA score through equations like

Pre-menopausal Predictive Index (PI) = $-12.0 + 2.38 \times \ln(HE4) + 0.0626 \times \ln(CA\ 125)$

Post-menopausal Predictive Index (PI) = $-8.09 + 1.04 \times \ln(HE4) + 0.732 \times \ln(CA\ 125)$

- Pre-menopausal: More than 13.1% ROMA score indicates a higher risk for ovarian cancer.

- Post-menopausal: More than 27.7% ROMA score indicates a higher risk for ovarian cancer.

### 3.9.3. Results Interpretation:

Using the obtained ROMA score result, the chatbot makes a specific interpretation:

For low-risk patients, it reassures the patient and refers them for follow-up.

It urges the patient to visit a health provider immediately and further testing for high-risk patients.

### 3.9.4. Personalized Recommendations:

Personalized Advice This would be offered in personalized advice according to the risk level of the patient. It covers what their next step, course of treatment, or lifestyle changes would be. For those at a high risk, the chatbot emphasizes the need for medical attention immediately.

### 3.9.5. Educational Support:

The chatbot offers necessary educational content as regards ovarian cancer, its symptoms, methods of diagnosis, and treatment options available. All this helps inform the patient on how to manage their health. The chatbot can even clarify frequent questions from patients ensuring that patients will have relevant information in time.

### 3.9.6. Interactive Feedback:

The chatbot responds dynamically to the most urgent patient query. It might explain to the patient the significance of ROMA score; answer questions about ovarian cancer; or even assist patients regarding how to get a second opinion or a specialist referral.

### 3.9.7. Accessibility and Engagement:

The chatbot is an open accessible platform available 24/7, hence any user can access information and get the required guidance at all times. This could be especially useful for continuous support for both the patients as well as the healthcare providers without necessarily requiring direct human-to-human interaction.

### 3.9.8. Diet Plans and Lifestyle Recommendations:

Another feature the chatbot contains is personalized diet and lifestyle advice. These are pre-existing clinical resources and expert guidelines for cancer patients. Such recommendations are extrapolated from books like Rebecca Katz's "The Cancer-Fighting Kitchen," which identifies the need for diet support in the management of cancer symptoms, as well as sides and treatments; "Eating Well Through Cancer" by Hollye Jacobs would also be suggested as a

list of foods that can be consumed in order to facilitate general wellness. End. It also encompasses general practices of wellness, hydration, stress management, and some very mild exercise that may also contribute to the patients' improvement on physical and emotional levels.

### 3.9.9. Hospitals and Medical Centres Information:

For those patients who might need more consultation on their part or further care, there is a list of nearby hospitals and medical centers of ovarian cancer treatment that the chatbot may provide. The information is cited from reliable health websites, such as Healthgrades and U.S. News & World Report, which update rankings of hospitals and reveal patient feedbacks. The chatbot may also suggest which one would best suit diagnosis, surgery, or treatment depending on the location of a patient. With this aspect, patients can easily seek expert care within their chosen institutions for ovarian cancer.

### 3.9.10. Patient Care and Support:

The chatbot offers emotional support and counselling on how a patient should care for himself or herself while going through the treatment procedures. Sources such as "Cancer Care: A Guide for Patients and Families" by Peter L. Greenwald may be useful in providing tips about common side effects, such as fatigue, nausea, and pain management. It could even provide suggestions of self-care practices that can help one maintain healthy psychology, such as mindfulness practice and mental health tips, among others.

A chatbot plays a key role in this ovarian cancer detection project bridging the gap between complex machine learning predictions and user-friendly patient engagement. Instead of merely helping to carry out and make sense of ROMA score, it also provides actionable insights and educational support. Such an integration would ensure that patients and clinicians have informed decision-making from reliable, real-time information with improved access, understandability, and tailored guidance. It enhances the ovarian cancer detection system and hence the general healthcare experience of users.

The diet plans, patient care guidelines, and hospital directories added to this chatbot can be used in order to provide a fully comprehensive tool for the medical professional as well as the patient, not only for short-term diagnostic support, but also for long-term health management. Patients may, then, enjoy much better general outcomes and wellness because they are continuously receiving support and actionable recommendations in accordance with their specific needs.

# CHAPTER 4

# 4. RESULTS

In this work, a machine learning-based diagnosis system, which predicts the risk of ovarian cancer, is proposed. XGBoost classifier has been used for classification, as it was better in performance compared to some traditional algorithms used for classification. Figure 3 compares the performance of various classifiers with an MCC score. Each of these models was trained independently on premenopausal and postmenopausal patient data. From the figure, it can be seen that KNN was the worst and ensemble algorithms such as Random Forest and AdaBoost are better. In each of the two classes, the best score for the performance was achieved by the XGBoost classifier and was therefore selected to be analyzed in more detail. Further to improve the accuracy of the classification, the Genetic Algorithm was utilized to determine the most representative set of features to be used by the XGBoost model and to improve overall performance.

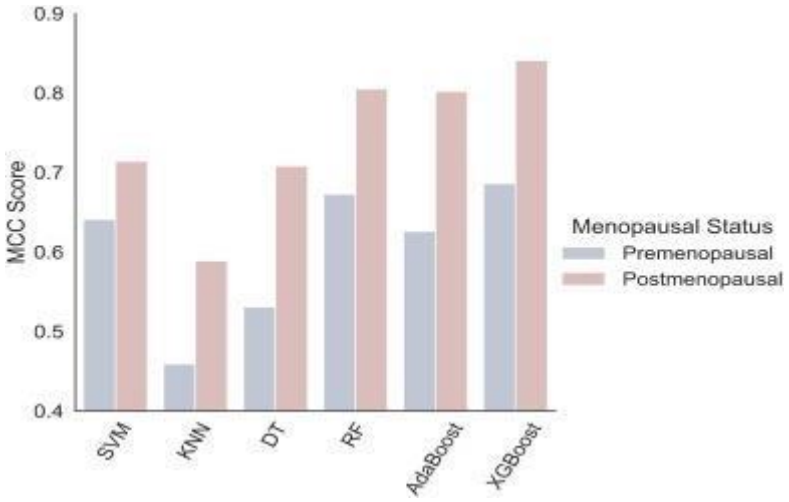



Fig. 3. Performance comparison of the XGBoost classifier with other classification algorithms

### 4.1. Performance of the XGBoost classifier:

The XGBoost classifier has a pretty good performance in terms of predicting the risk of ovarian cancer (OC) for premenopausal and postmenopausal women. Different feature selections carried out through the Genetic Algorithm for each group picked out different sets of highly

significant biomarkers, inferring the difference of OC predictors among premenopausal and postmenopausal patients.

Table 4 gives the selected features for each group. Six features for premenopausal and seven for postmenopausal were selected as the best models for prediction. The features thus optimized by GA improved the accuracy of the classifier and provided an optimized approach for assessing the risk of OC.

Although the GA would proficiently reduce the most significant predictors, this scheme is not helpful to reveal the degree and the type of contribution from each feature to the prediction outcome based only on the XGBoost model. We thus combined SHAP in our analysis to make sense of the predictions of the model and how individual features influence the chance of OC. The combination of XGBoost with both GA and SHAP enables the attainment of high-performance combinations while the interpretability is at its best to meet the clinical demands.

| **Approach** | **Menopause Status** | **Accuracy** | **Sensitivity** | **Specificity** | **Precision** | **G-mean** | **ROC-AUC** | **MCC** |
|---|---|---|---|---|---|---|---|---|
| XGBoost | Premenopausal | 0.861 | 0.712 | 0.935 | 0.861 | 0.808 | 0.824 | 0.686 |
| | Postmenopausal | 0.949 | 0.979 | 0.833 | 0.962 | 0.889 | 0.906 | 0.841 |
| | Total Average | 0.905 | 0.845 | 0.884 | 0.912 | 0.849 | 0.865 | 0.764 |
| XGBoost + GA | Premenopausal | 0.896 | 0.773 | 0.955 | 0.889 | 0.857 | 0.864 | 0.757 |
| | Postmenopausal | 0.958 | 0.989 | 0.833 | 0.963 | 0.893 | 0.911 | 0.868 |
| | Total Average | 0.927 | 0.881 | 0.894 | 0.926 | 0.875 | 0.887 | 0.813 |

Table 3. Performance measures obtained from the proposed approach.

### 4.2. Performance after incorporating feature selection strategy

Identifying principal risk factors is critical for proper OC diagnosis. When lots of features are used, it complicates the decision-support system with an enormous amount of testing associated, adding cost, and introducing bias and the problem of overfitting. In order to reduce the above problems, we used a Genetic Algorithm-based Feature Selection (GA-FS) technique to optimize feature selection.

In this step, the reduced feature set obtained was given to the model for further training to achieve better performance. The results are given below in Table 3; it has been seen that 89.6% accuracy was obtained for premenopausal patients and ROC-AUC score of the model increases up to 0.864 which signifies a balanced and robust classifier with the discrimination of pre- and post-menopausal patients. For the patients postmenopausal, the attained sensitivity score reached 98.9%, which nearly perfectly classified patients at risk of OC, a critical property in clinical diagnostics. Fig. 4 shows model performance after feature selection.

GA identified different subsets of features for premenopausal and postmenopausal patients, which means that the groups varied in risk factors. Table 4 Summarizes the chosen features: six features are the most important for prediction in premenopausal women, while seven features are important for postmenopausal women. However, while these chosen features improve the performance of the model, GA alone does not tell how a feature contributes to OC prediction.

To surpass this limitation, the SHAP (SHapley Additive exPlanations) approach was applied. This ensured that a more in-depth understanding of the influence an individual feature had on the predictions of the model was understood. This approach achieved global and local interpretability, ensuring the correctness of the developed diagnostic model also translates to transparency in informing more aware decisions in clinical applications.

| **Menopausal Status** | **Features** | | | | | | |
|---|---|---|---|---|---|---|---|
| Premenopausal | CA125 | HE4 | CEA | ALB | TP | LYM | |
| Postmenopausal | CA125 | CA199 | RBC | HCT | K | MONO % | Na |

Table 4. Features selected by the GA.

# CHAPTER 5

## 5. DISCUSSION

The paper introduces a new, interpretable ML-based diagnostic system for predicting the risk of ovarian cancer (OC), with special application to premenopausal and postmenopausal women. The genetic algorithm suggested a reduced subset of the most relevant critical risk factors from an initial set of 48 features, which greatly simplified the model and also showed the importance of features to differ with menopausal status. The use of the XAI techniques, especially the SHAP method, has been used to provide interpretability by indicating how each feature contributes to the prediction of OC. For premenopausal, the most important predictors were CEA, HE4, and ALB, whereas, for postmenopausal patients, the most influential biomarker was CA125. Furthermore, SHAP values also point to the feature that increased the risk: in the case of a premenopausal woman, a high positive value of HE4, CEA, and TP increased it, whereas ALB levels seemed to reduce it. The combination of GA with SHAP is a step closer to providing a clinically relevant model: fewer features, with higher accuracy would imply a more cost-effective, transparent and reliable approach for diagnosing the risk assessment of OC.

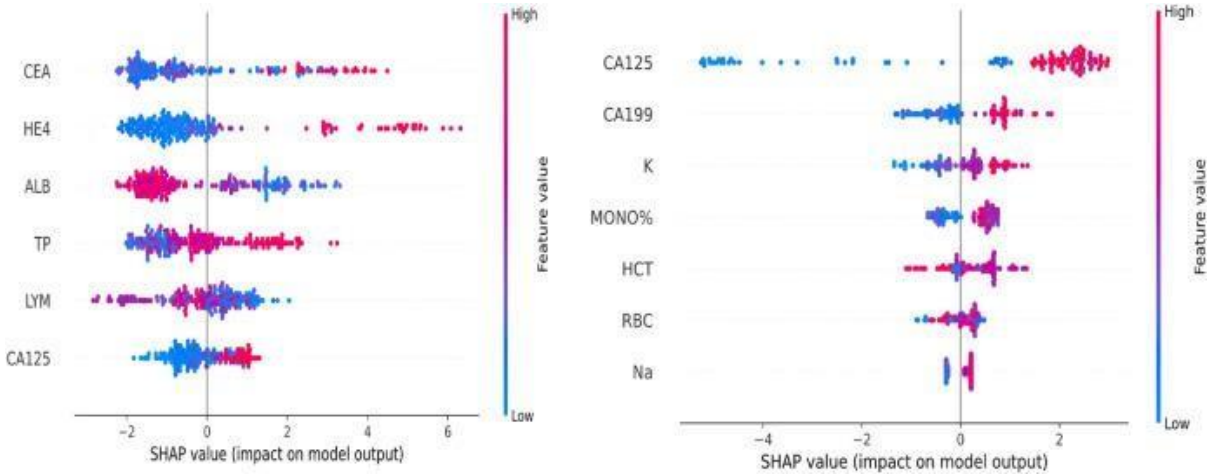



Fig. 4. Relationship of the predictor variables with the target in the - (a) premenopausal population and (b) postmenopausal population.

### 5.1. Analysis of the risk factors identified by the GA in the diagnosis of OC

Only similar risk factors were seen by the GA for the diagnosis of ovarian cancer, particularly CA125, both in the pre- and postmenopausal groups. All the other risk factors and diagnostic tools were inconsistent with both the groups. The tool applied for OC screening is very well known to be a common biomarker: CA125. Cancer Antigen 125 has the protein levels measured in the blood; high results indicate the risk of ovarian cancer. This is supported by SHAP analysis (Fig. 7), where high CA125 levels are associated with an elevated risk of cancer. HE4 is another important biomarker for OC and has been significantly influenced by menopausal status. In this study, relevance of HE4 was only observed in the case of premenopausal patients and is consistent with previous studies that showed the diagnostic utility of HE4 only in the case of premenopausal females. Similar to several other studies, HE4 has also been reported to be of little utility in postmenopausal OC diagnosis, and hence this study's results are further validated. Several studies have applied both CA125 and HE4 in ROMA-algorithm that is the most current risk assessment test for OC; however, ROMA's sensitivity and specificity primarily for the premenopausal patients are still limited (Table 5). For this reason, an ongoing approach is to identify new biomarkers that may provide an important aid in enhancing the diagnosis of OC.

In our experiments, we discovered even more biomarkers that may possibly predict diagnosis of OC. Serum low levels of albumin (ALB), associated with hypoalbuminemia common in patients with cancer who are malnourished and have lost weight, have been reported as indicative of a higher risk of OC (Fig. 7). Additional studies also suggested that preoperative serum ALB levels may serve as an independent survival predictor for epithelial OC patients. For postmenopausal patients, CA19-9 was selected as the most important predictor following CA125, and elevated serum CA19-9 level was also strongly associated with a higher risk of cancer (Fig. 7b). These have been in agreements with other studies where values of CA19-9, CA-125, and CEA have appeared to be good predictors in the prediction of levels of OC.

Our study has given rise to the emergence of CEA as a pertinent predictive marker among premenopausal patients whose CA19-9 (>39 U/mL) or values of CEA (>3.8 U/mL) gave a good positive association with malignancy of OC.

Besides, these are MONO%, HCT, and K that appeared to become pertinent predictors in postmenopausal women and, on the other hand, for premenopausal, TP and LYM. These variables are in line with the studies that are relevant to the diagnosis and prognosis for OC

patients. Properties like Na and RBC showed minimal contribution as indicated by the SHAP values. These features may be there just due to the stochastic nature of the machine learning and feature selection technique and slight differences appear in the chosen features while tweaking the GA parameters or iteration counts. However, most critical predictors did not change. Adding SHAP to evaluate and emphasize the risk factors that are most influential in GA's chosen feature subset provides a more robust method for the proper diagnosis of OCs.

### 5.2. Interpretation of the model using XAI (SHAP)

Explainable AI (XAI) introduces transparency to model prediction and enables tracking the outcome results, thereby enhancing the analysis of model outcomes and the explanation behind predictions. The feature importance plots in Fig. 4 describe the contribution of various features toward the prediction of OC and highlight whether each has a positive or negative association with the target variable. Partial dependence plots are used to further analyze the relations between biomarkers and cancer risk with the purpose of determining threshold points (Fig. 5 and Fig. 6). These plots are based on Shapley values and indicate the way changes in feature values affect the risk; therefore, they may detect linear or monotonic relationships but also complex ones, between features and targets. For example, increasing TP values almost-linearly increase cancer risk; Fig. 8. Furthermore, by design, each plot is populated with the interaction that is most variable with the feature of interest, for example, ALB and CEA both related to HE4 (Fig. 5). Change in Shapley values reflect abrupt changes in risk: for example, while CA125 has negative Shapley values below 30 U/mL in postmenopausal women, a sharp increase in risk occurs for values above this threshold (Fig. 6). Other risk factors show similar thresholds, but their threshold values are also not absolute. They depend on data and can vary from one dataset to the other.

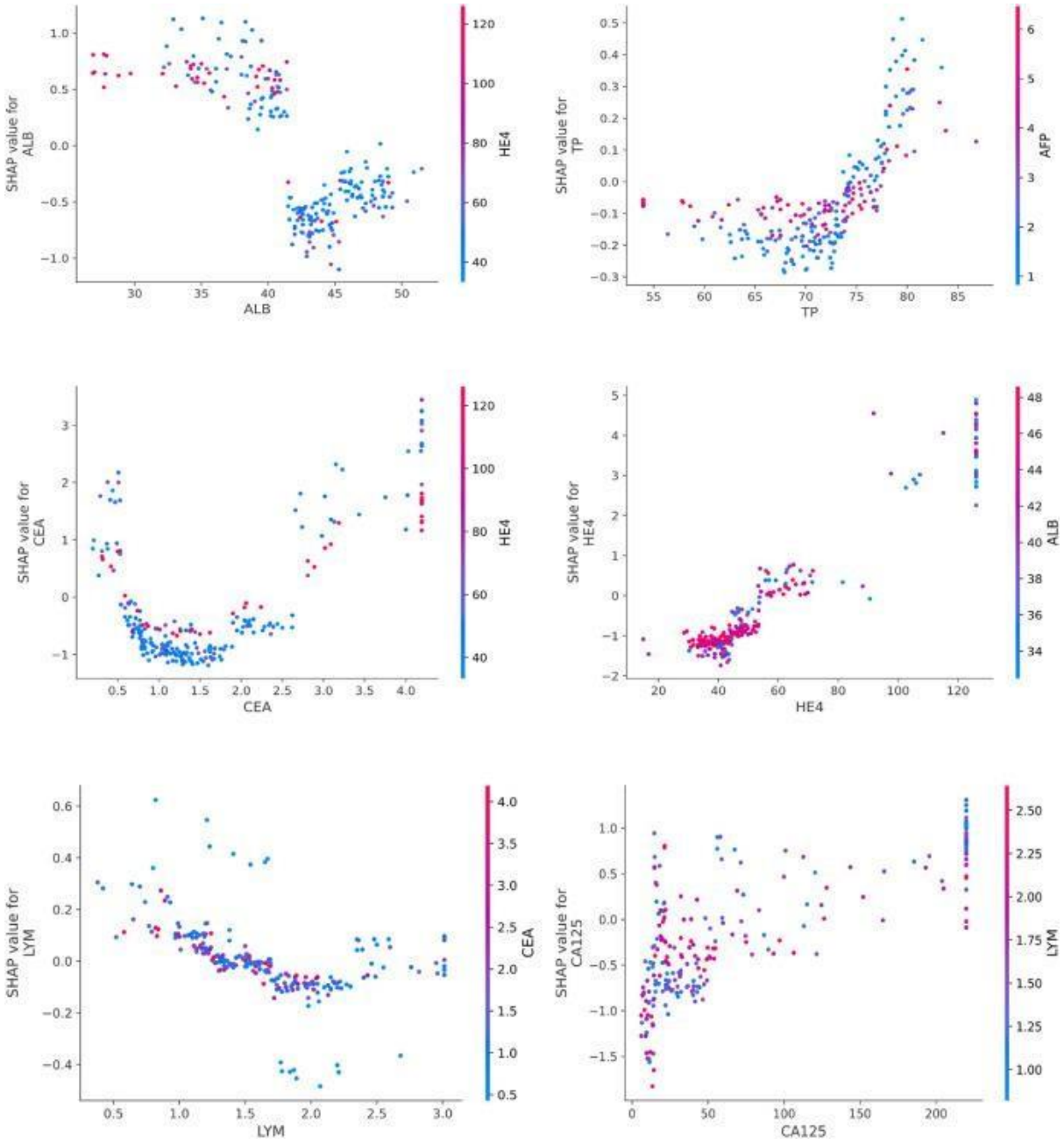



Fig. 5. Partial dependence plots for risk factors in premenopausal patients.

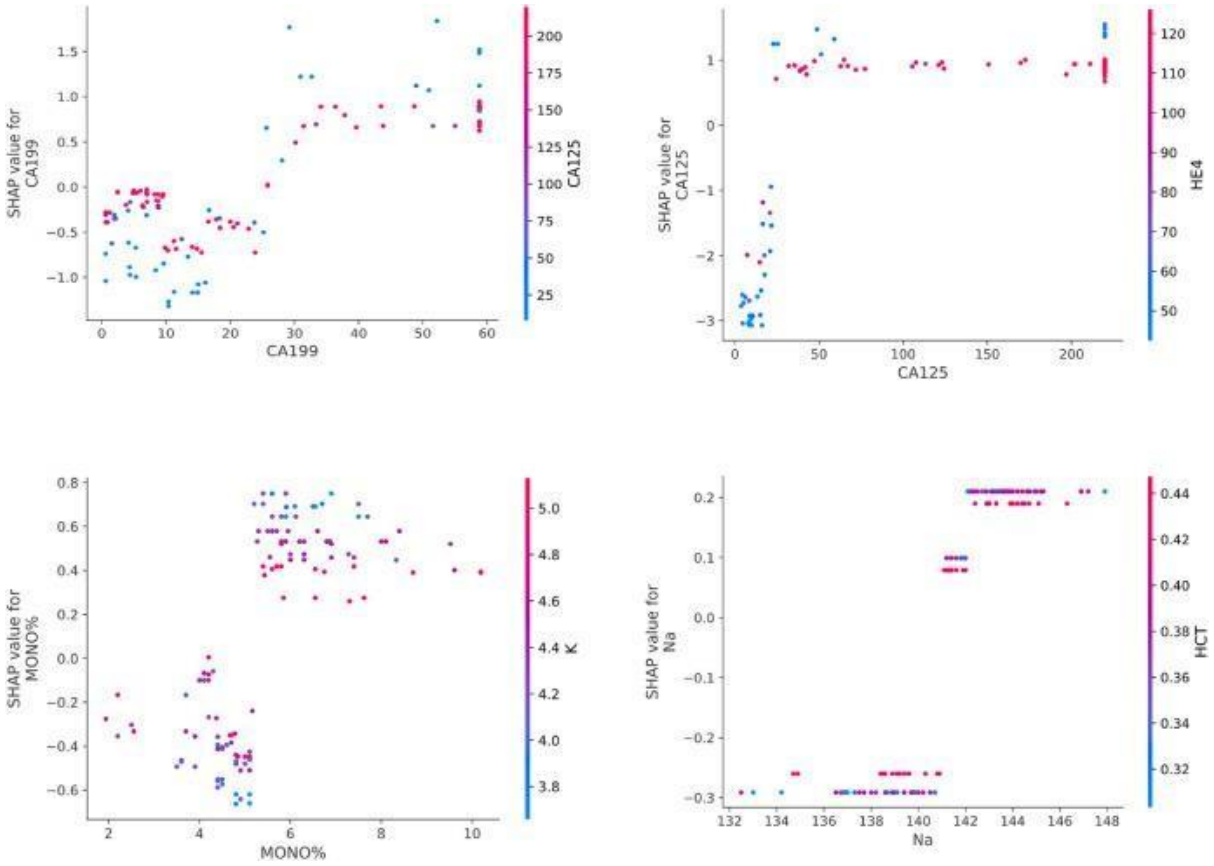



Fig. 6. Partial dependence plots for risk factors in postmenopausal patients.

### 5.3. Local interpretability using SHAP

SHAP provides an explanation of individual predictions by the ML classifier to the clinician as to what drives each prediction.


Fig.7. SHAP force plot for the interpretation of individual predictions.

In Fig. 7, SHAP explanations are applied to better represent model predictions for two patients. Fig. 7a shows a premenopausal woman with high-risk cancer, where the SHAP force plot

reflects how HE4, followed by ALB and CA125, contributes positively, pushes the prediction from the baseline of 0.26 up to 0.9. HE4, high, and ALB is very low, both are risk factors for ovarian cancer (OC) match Fig. 5. Fig. 10b translates a low probability of risk for a postmenopausal female where CA125 acts as the primary source of pulling the probability toward zero, indicating that even low levels of CA125 decrease the risk of OC, as seen in Fig. 6.


Fig. 8. Interpretation of wrong predictions made by the model using SHAP force plot.

SHAP force plots are particularly useful in analyzing model misclassifications, as illustrated by the two false-positive cases depicted in Fig. 5. In Fig. 5a, a premenopausal patient with low OC risk was predicted to be positive with probability 0.58. Even low levels of HE4, CEA, and TP theoretically would suggest lower cancer risk, but it is difficult to extract a clear sense from this because of the near-boundary level of ALB (41.4) and a very unusual LYM ratio (1.09), shifting the prediction rightward. This makes evident that even slight deviations in the level of biomarkers under discussion may overturn the predictions, thus again requiring close clinician review. Such abnormal values of data, as obtained for the LYM ratio, may also result due to natural variation as well as possible entry errors in data; hence, further manual scrutiny at times is recommended. Thus, SHAP force plots are a highly useful tool for validating model predictions and helping clinicians reach an informed decision.

# CHAPTER 6

## 6. CHATBOT INTEGRATION IN OVARIAN CANCER DETECTION

The ovarian cancer diagnostic system includes a chatbot: it is an interactive interface to the model of machine learning, and an effective virtual assistant for patients and clinicians in the process of diagnosis. The primary objective of the chatbot will be to collect relevant input from patients, including biomarker values, age, and menopausal status, and then forward it to the machine learning model for analysis. It will thereby ensure that the generated system comes up with accurate calculations of the risk of cancer under the latest accessible data.

It allows users to input the data, generate risk predictions, and explain these findings in simple terms. Moreover, it gives transparency to which biomarkers contributed the most to the risk assessment. That would be a significant feature particularly in healthcare scenarios where patients and clinicians often need to understand why a particular medical decision has been made. It bridges the gap between two extremes: medical jargon and user-friendly insights that could make non-experts understand how individual risk factors influence the diagnosis. Besides this, the AI model will aid the clinician with decisions based on suggestions of possibly further courses of action through additional diagnostic tests or treatments, informs patients regarding lifestyle recommendations, follow-up tests, and information about regional healthcare facilities, all based on the patient's needs according to his individual biomarkers.

For example, if the patient is premenopausal and her risk score is high based on elevated biomarkers, for example, HE4 or CA125 levels, she will be given information on such values by the chatbot and encouraged to go see a consultant without wasting any further time.

If a postmenopausal woman's risk happens to be low on account of normal biomarker levels, then the chatbot will reassure her and give her preventive advice regarding regular monitoring. It will help to keep the patient in control of their health and encourage them to be proactive in its management. This would be an excellent source for patients to inquire further, get a second opinion, or have their medical terms explained concerning their diagnosis. Since there is a chatbot, this makes the patient communication accessible and available continuously 24/7 as opposed to regular working hours, thus expanding patient engagement and education.

It increases the incorporation of such a chatbot in the machine learning diagnostic system, giving clinicians more credible information, while patients receive a more customized experience. Such an approach ensures that deployment of AI not only goes to predict the probability of disease but also clarify the underlying reasons for those predictions in an easier digestible format that enhances trust in the technology.

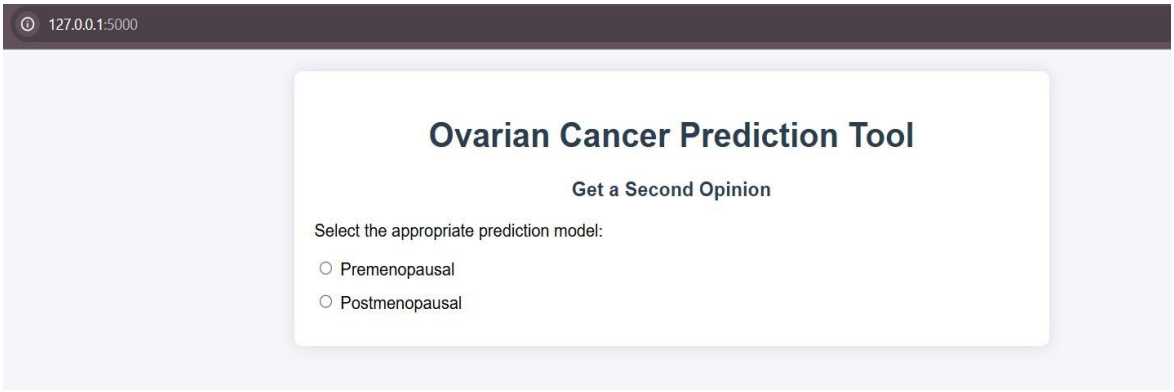



Fig.9. Interface to select the menopause status

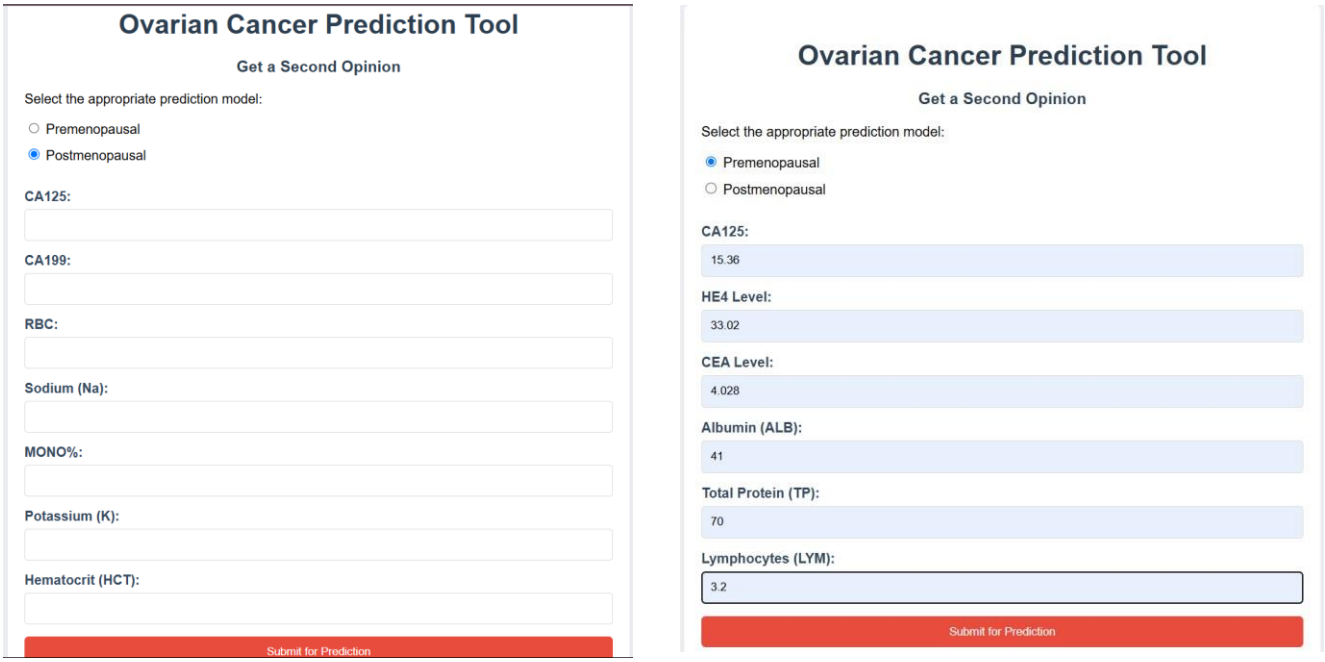
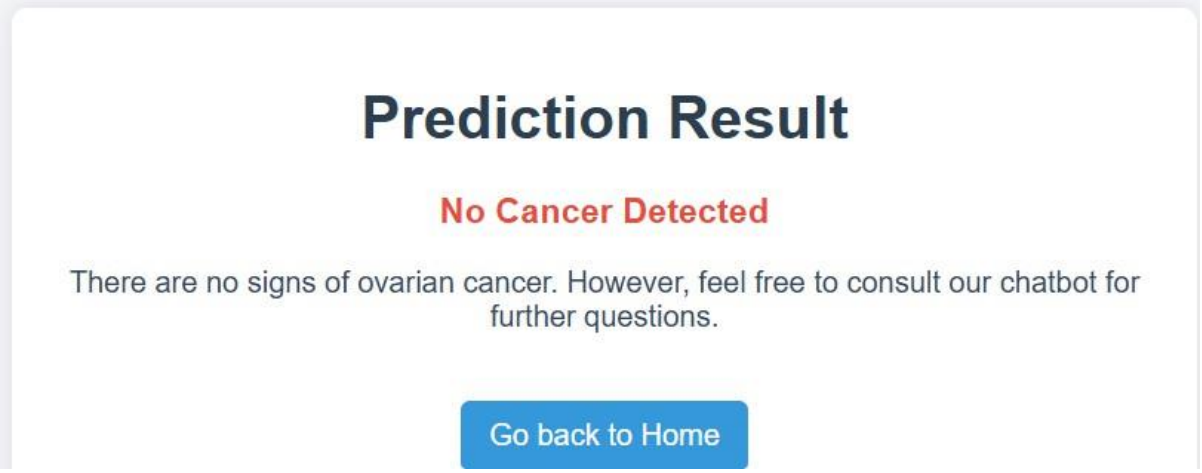
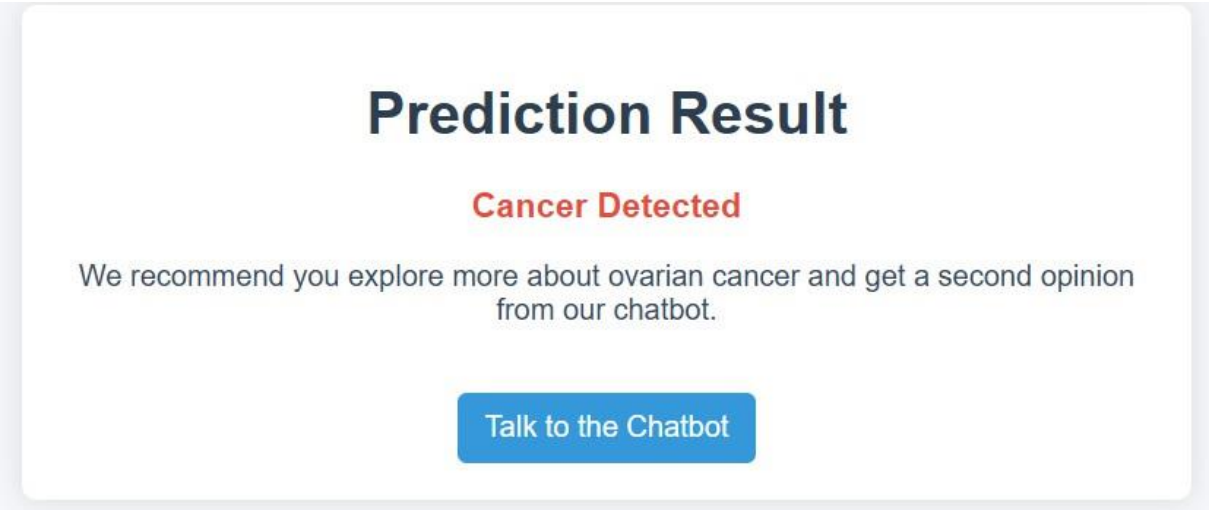
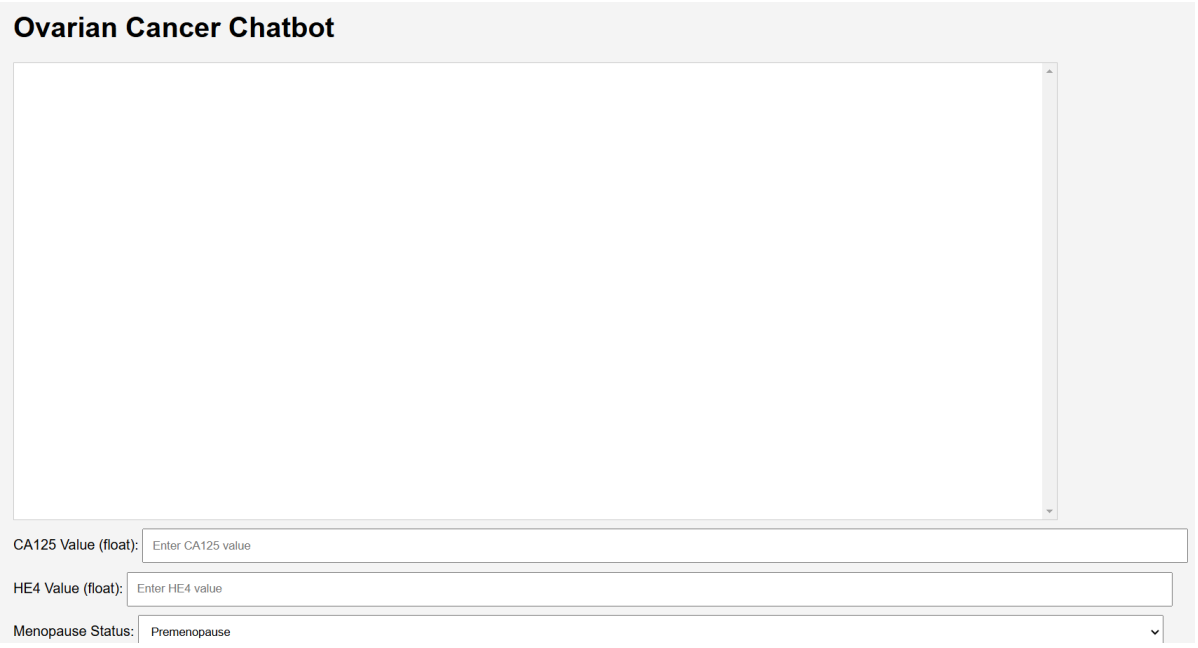
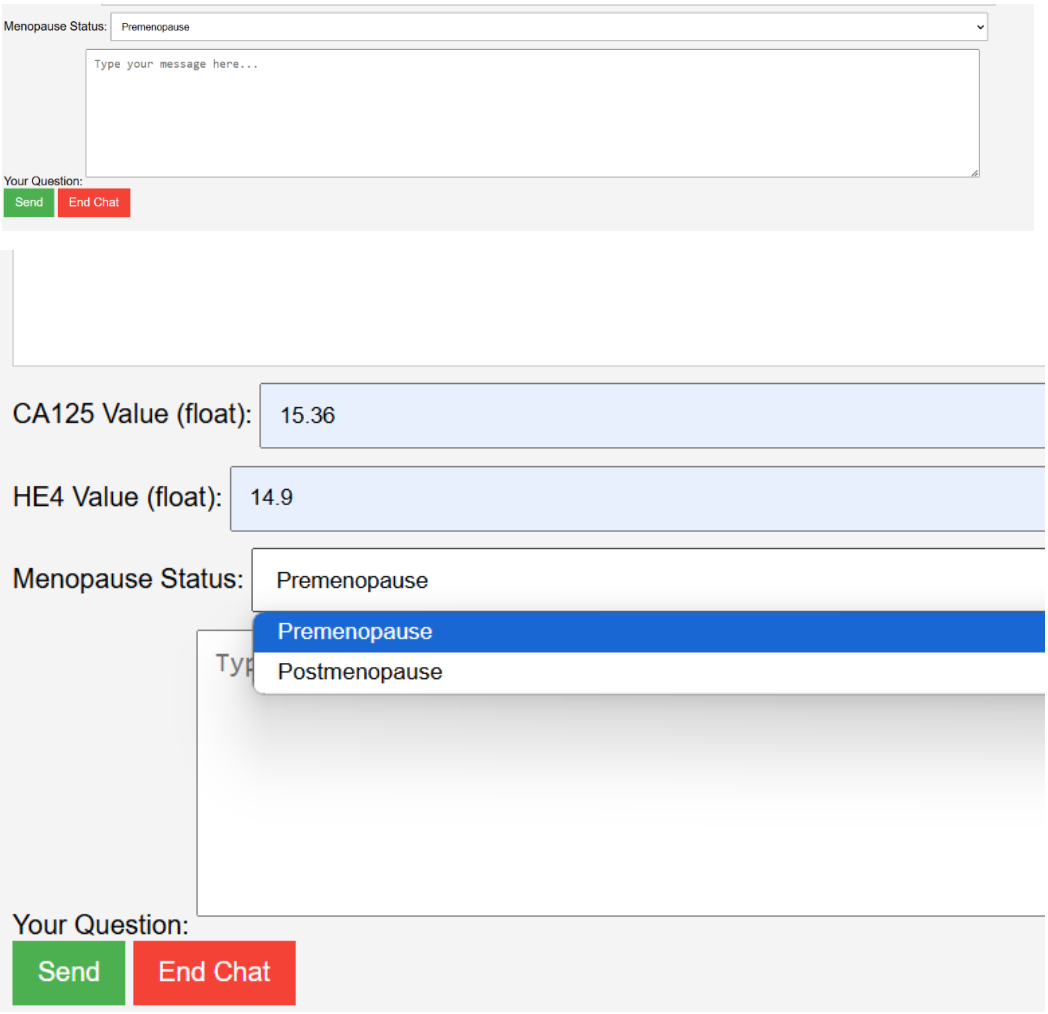



Fig.10. Giving the features value from the menopausal status a) Postmenopausal b) Premenopausal


Prediction Result
No Cancer Detected
There are no signs of ovarian cancer. However, feel free to consult our chatbot for further questions.
Go back to Home


Prediction Result
Cancer Detected
We recommend you explore more about ovarian cancer and get a second opinion from our chatbot.
Talk to the Chatbot


Ovarian Cancer Chatbot
CA125 Value (float):
Enter CA125 value
HE4 Value (float):
Enter HE4 value
Menopause Status:
Premenopause



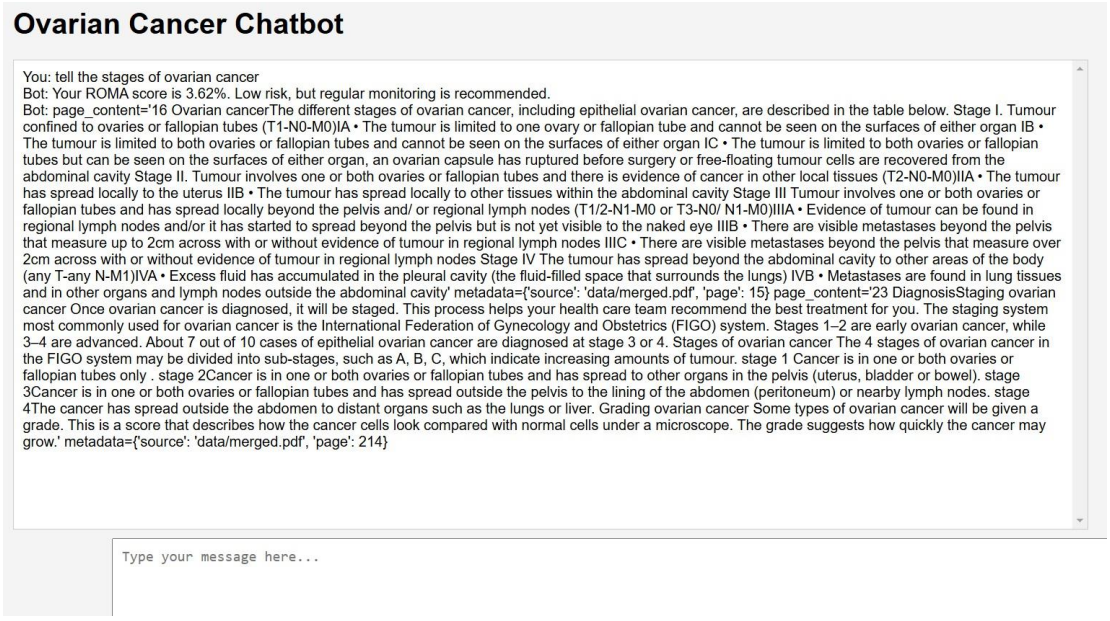




Fig.11. Output of the user given data

# CHAPTER 7

# 7. CONCLUSION

Ovarian cancer is one of the main causes of death for women from the disease, and the early detection of this cancer remains the only way to increase the chances for survival. The biomarkers are not reliable and often result in the detection of this type of cancer only in later stages, which leads to more reasons for developing even better diagnostic tools. The project of ours focused on the machine learning-based diagnostic system designed to identify ovarian cancer with higher accuracy. Our methodology employs explainable AI, to achieve better clinician support through transparency and deep insights into what models are predicting.

It was, therefore, possible to identify relevant biomarkers for both pre- and postmenopausal women through genetic algorithms along with SHAP data-driven feature selection at high precision. While CA125 is a crucial predictor for both, other markers like HE4 apply uniquely for premenopausal and CA19-9 for postmenopausal women. The features selected were fed into an XGBoost model, which achieved a high ROC-AUC score on the pre- and postmenopausal populations at 0.864 and 0.911, respectively, significantly outperforming traditional methods of diagnosis.

This allowed interpretation of the predictions by the models, identifying the critical biomarkers with relevant threshold values- CEA, HE4, and ALB for premenopausal women and CA125 and CA19-9 for postmenopausal women. From this, new avenues in the diagnosis of ovarian cancer have been unveiled as well as promising biomarker candidate's worth further clinical research. We will gather more data from different demographics to reduce bias and enhance robustness in light of the data limitations.

Worked on a graphical user interface that will allow clinicians to see the individual predictions visually and understand what features are contributing to the model's decisions. The user-friendly interface is an important tool for the most reliable diagnosis and every individual patient assessment so that the diagnostic process becomes transparent and trustworthy. Outcomes of our project show that a hybrid AI framework with XAI integration could potentially deliver a reliable, interpretable, and impactful tool for early ovarian cancer diagnosis that, in turn, supports timely and informed decision-making in practice.

## Repository and Code:

The associated codes, results, and other supplementary files are provided in the following repository: GitHub Repository.

# REFERENCES


[ 1 ] Asif Newaz, Abdullah Taharat, Md Sakibul Islam, Khairum Islam, A.G.M. Fuad Hasan Akanda, An ML-based decision support system for reliable diagnosis of ovarian cancer by leveraging explainable AI, Informatics in Medicine Unlocked, Volume 49, 2024, 101553, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2024.101553.

[ 2 ] Cooper, Geoffrey M., and Robert E. Hausman. "The development and causes of cancer." *The cell: A molecular approach* 2 (2000): 725-766.

[ 3 ] Siegel, Rebecca L., Kimberly D. Miller, Hannah E. Fuchs, and Ahmedin Jemal. "Cancer statistics, 2022." *CA: a cancer journal for clinicians* 72, no. 1 (2022): 7-33.

[ 4 ] Bray, Freddie, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." *CA: a cancer journal for clinicians* 68, no. 6 (2018): 394-424.

[ 5 ] Torre, Lindsey A., Britton Trabert, Carol E. DeSantis, Kimberly D. Miller, Goli Samimi, Carolyn D. Runowicz, Mia M. Gaudet, Ahmedin Jemal, and Rebecca L. Siegel. "Ovarian cancer statistics, 2018." *CA: a cancer journal for clinicians* 68, no. 4 (2018): 284-296.

[ 6 ] Granberg, Seth, Matts Wikland, and Inge Jansson. "Macroscopic characterization of ovarian tumors and the relation to the histological diagnosis: criteria to be used for ultrasound evaluation." *Gynecologic oncology* 35, no. 2 (1989): 139-144.

[ 7 ] Aslan, Koray, M. Anıl Onan, Canan Yilmaz, Neslihan Bukan, and Mehmet Erdem. "Comparison of HE 4, CA 125, ROMA score and ultrasound score in the differential diagnosis of ovarian masses." *Journal of Gynecology Obstetrics and Human Reproduction* 49, no. 5 (2020): 101713.

[ 8 ] Jacobs, I., D. Oram, J. Fairbanks, J. Turner, C. Frost, and J. G. Grudzinskas. "A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer." *BJOG: An International Journal of Obstetrics & Gynaecology* 97, no. 10 (1990): 922-929.

[ 9 ] Miotto, Riccardo, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T. Dudley. "Deep learning for healthcare: review, opportunities and challenges." *Briefings in bioinformatics* 19, no. 6 (2018): 1236-1246.

[ 10 ] Kleppe, Andreas, Ole-Johan Skrede, Sepp De Raedt, Knut Liestøl, David J. Kerr, and

Håvard E. Danielsen. "Designing deep learning studies in cancer diagnostics." *Nature Reviews Cancer* 21, no. 3 (2021): 199-211.

[ 11 ] Newaz, Asif, Sabiq Muhtadi, and Farhan Shahriyar Haq. "An intelligent decision support system for the accurate diagnosis of cervical cancer." *Knowledge-Based Systems* 245 (2022): 108634.

[ 12 ] Peiffer-Smadja, Nathan, Timothy Miles Rawson, Raheelah Ahmad, Albert Buchard, P. Georgiou, F-X. Lescure, Gabriel Birgand, and Alison Helen Holmes. "Machine learning for clinical decision support in infectious diseases: a narrative review of current applications." *Clinical Microbiology and Infection* 26, no. 5 (2020): 584-595.

[ 13 ] Mi, Qi, Znati Jingting, Fan Ty, Li Zhenjiang, Xu Jundong, Chen Bin, Zheng Lujun, and Lu Xiao. "Data for: Using machine learning to predict ovarian cancer." *Mendeley Data, Version* 11 (2020).

[ 14 ] Atallah, Ghofraan Abdulsalam, Nor Haslinda Abd. Aziz, Chew Kah Teik, Mohamad Nasir Shafiee, and Nirmala Chandralega Kampan. "New predictive biomarkers for ovarian cancer." *Diagnostics* 11, no. 3 (2021): 465.

[ 15 ] Atallah, Ghofraan Abdulsalam, Nor Haslinda Abd. Aziz, Chew Kah Teik, Mohamad Nasir Shafiee, and Nirmala Chandralega Kampan. "New predictive biomarkers for ovarian cancer." *Diagnostics* 11, no. 3 (2021): 465.