Machine Learning Strategy for Ovarian Cancer Diagnosis

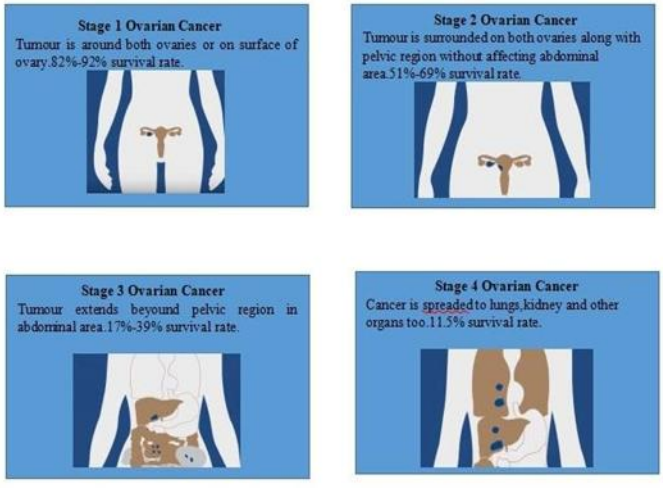
Sneha Unnarkar[1], Bijal Patel[2]
[1,2]*Computer Engineering Department*
SAL Education, Ahmedabad, Gujarat, India. Gujarat, India.

***Abstract*—Ovarian cancer is one of the leading causes of cancer-death in female, if it is not detected at an early stage. Surgery, ancestry testing, ultrasound, CT-Scan, blood test like CA125 and extra medical examination are the main methods used to diagnose ovarian tumors whether it is benign or malignant in women. As timely detection of cancer is an important aspect, Machine learning is an emerging field that can make accurate projections by making inferences on data and may play a crucial role in Ovarian Cancer Prediction. According to observations, there are various Machine Learning Algorithms such as Support Vector Machine, K- nearest neighbor and Logistic regression etc. that may help to prevent cancer death if it is diagnosed at an early stage. The objective of the current study is to diagnose ovarian cancer accuracy using different Machine learning strategies.**

***Index Terms*—Ovarian Cancer, Early stage, Medical Examination, Machine learning algorithms.**

I. INTRODUCTION

According to the World Health Organization, cancer is a major worldwide health issue leading to fatalities. Cancer is characterized as a collection of cells originating from specific regions of the human body, which frequently spread to distant metastatic locations. Ovarian cancer (OC) is a specific type of cancer originating in the ovaries, characterized by the rapid multiplication of abnormal cells that form tumors [1]. Epithelial ovarian cancer (EOC) accounts for around 90% of ovarian cancer cases. Ovarian cancer has various cellular origins. The term "tubo-ovarian cancer" is commonly used to describe the disease because it can present as a tumor in the ovary, Fallopian tube, or primary peritoneal region. Type I tumors, including low-grade serous, mucinous, endometrioid, and clear cell tumors, have a tendency to grow slowly and are less aggressive, making them easier to detect in the early stages. On the other hand, type II tumors, such as high-grade serous carcinomas (HGSC), undifferentiated carcinomas, and carcinosarcomas, are more aggressive and can arise from the epithelium of the Fallopian tubes and/or the surface of the
Ovaries [9]. Certain tumors, such as serous cystadenomas, originate from the epithelial surface of the ovaries. These cystadenomas are often found on both ovaries and consist of a watery fluid. In contrast, mucinous cystadenomas are composed of multiple cells and contain a fluid resembling mucus. Additionally, there are ovarian germ cell tumors like mature cystic teratomas, commonly referred to as dermoid cysts [5]. Among young women, these tumors are the most common in the ovaries and usually consist of a combination of mature tissues derived from two or three different germ cell layers. Similar to benign tumors, they arise from the surface epithelium of the ovaries and exhibit similarities to serous or mucinous cystadenocarcinomas. When it comes to the four stages of ovarian cancer, the maximum survival rate for a woman is five years. Different machine learning models are employed to forecast parameters like precision, recall, and F1- score that helps in better prediction [10].



Cervical examination, colour Doppler ultrasound (USG), serum CA125, CT, transvaginal sonography (TVS), trans-abdominal USG, and magnetic resonance imaging are the most frequently used screening techniques for ovarian cancer [10]. In the medical field, machine learning can be implemented for analyzing patient data, and also for diagnosing diseases. The machine takes the important features of the patient as input and produces an accurate diagnosis as output. Many different Machine Learning algorithms are used on different repositories to get some solutions like early-stage cancer detection with near to exactness accuracy.

II. LITERATURE REVIEW

It provides a concise overview of previous studies pertaining to the research subject, aiming to establish the connection between current understanding and novel discoveries.

In this study [14], the author utilized decision tree classifiers to predict ovarian cancer survival based on factors such as tumor size, age, tumor mobility, and surface characteristics. Various machine learning algorithms, including Logistic Model, Decision Tree, and Multi-layer Perceptron, were tested. Cross-validation was employed to assess the model's performance.

In this paper [15], Byoung-Gie Kim proposes the development of a Gradient Boosting Algorithm, which is expected to outperform other algorithms. The dataset used for this study was obtained from Samsung Medical College and Asan Medical Center. The findings could potentially assist patients with epithelial ovarian cancer in selecting optimal treatment options and improving individual outcome prediction.

In this mentioned work [16] Mingyang Lu suggests using the Maximum Redundancy and Minimum Relevance feature selection method on the dataset to find the most pertinent characteristics for building a straightforward decision tree in their research. The goal of the study is to maximize performance using the two biomarkers HE4 and CEA. The results imply that CEA is a useful diagnostic for ovarian cancer prediction in patients with low HE4 levels. This finding creates new avenues for examining CEA's function in ovarian cancer research.

The researcher [17] has used and develop The Clinical Proteomic Tumour Analysis Consortium (CPTAC) data site has the datasets used in this investigation. (https://cptac-data- portal. georgetown.edu/). It has also used For the risk assessment of ovarian cancer in individuals with a pelvic mass, a number of biomarkers have been employed, including CA12511, HE412 and osteopontin13. Here, we go over how the Decision Support System was created using a machine learning pipeline. Proteomic dataset explainable dss, feature selection with correlation analysis feature selection with relief

III. METHOD LEARNING DECISION TREE.

In their paper [18], the author discusses the application of a univariate statistical method to text data by utilizing bootstrap resampling with Bag of Words. Feature extraction was conducted using Principal Component Analysis and Multiple Correspondence Analysis. The dataset was obtained from Hospital HM Sanchinarro, Spain. However, while univariate analytic techniques provide valuable information, they have the limitation of disregarding inter dependencies between variables.

Several approaches have been employed to tackle the challenge of accurate predictions in the context of ovarian cancer. The mentioned paper [19] discusses the utilization of multiple ensemble machine learning algorithms for predicting ovarian cancer across various datasets. The author of the study

conducted data preprocessing and employed statistical and machine learning techniques to identify important features for the early diagnosis of ovarian cancer patients. This section investigates a range of Machine Learning techniques and the utilization of multiple datasets. Interestingly, smaller datasets have shown promising accuracy results across various approaches. In light of this, a novel machine-learning model has been developed, incorporating the many algorithms such as K-means clustering algorithm, Decision Tree, and logistic regressor. This combined approach aims to enhance accuracy predictions in the field of ovarian cancer research.

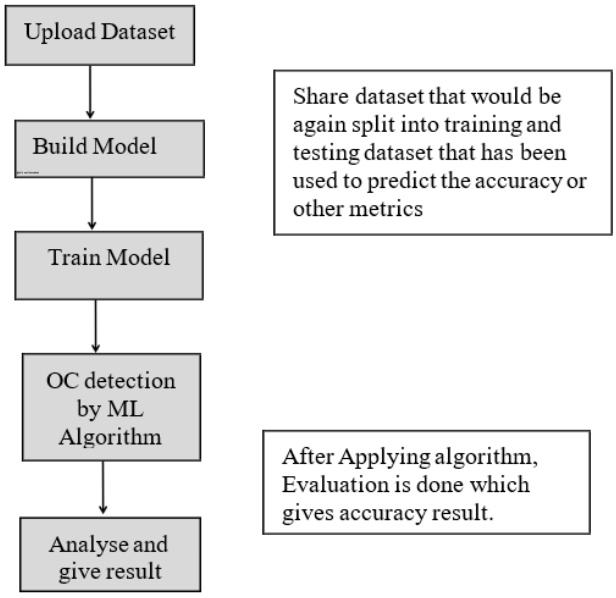


Figure 1 Basic Machine Learning Generalized Process

IV. PROPOSED METHODOLOGY

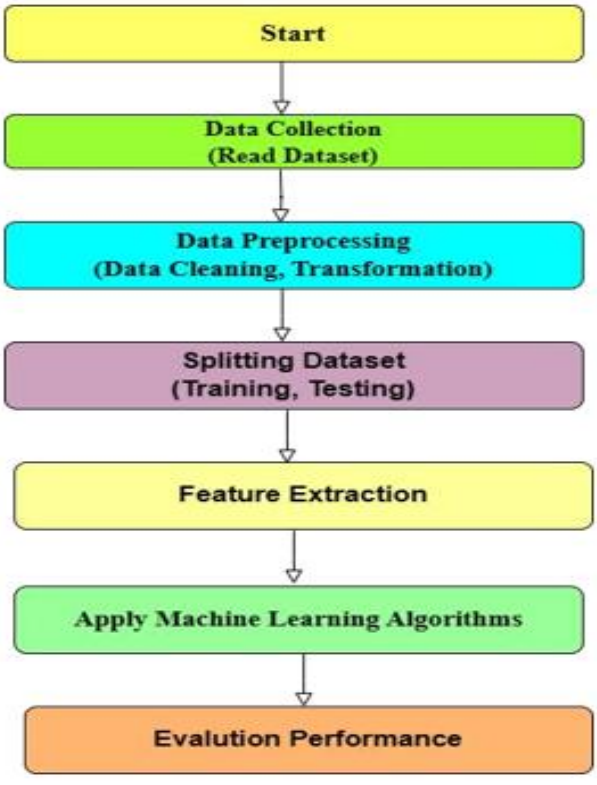


Figure 2 Proposed Workflow

The methodology section provides a detailed description of the research design, data collection methods, sample selection, data analysis techniques, and any tools or instruments utilized in the study. It serves as a guide for other researchers to understand and potentially replicate the study, ensuring transparency and rigor in the research process.

A well-defined and robust methodology enhances the credibility and validity of the research findings. The research paper utilizes a dataset to conduct the study and gather relevant data. The GEO database refers to the Gene Expression Omnibus database, which is a public repository for gene expression data. It is maintained by the National Center for Biotechnology Information (NCBI) and provides researchers with access to a vast collection of gene expression data from various organisms and experimental conditions.

The dataset that has been used with our proposed system is available on Kaggle, GSE38666.csv and Feature.csv. First of all, the code imports Overall, the code imports the required libraries and modules for performing data analysis, clustering, and evaluation tasks. The next step is to in summary; the code reads the data from a CSV file named 'GSE38666.csv' located at '/content' and stores it in a pandas Data Frame named 'data'. It accomplishes this by directly providing the file path to pd.read_csv() or by first opening the file with open () and then passing the file object to pd.read_csv(). Then we perform transposing of data and the reason behind transpose is certain machine learning algorithms or libraries may expect input data to be in a specific format, such as having samples (observations) as rows and features (variables) as columns. Transposing the data helps conform to the expected format if the original data has a different orientation. The code performs k-means clustering on the tcga_input data, retrieves cluster centers and labels, and calculates evaluation metrics (silhouette score and Davies- Bouldin index) to assess the quality of the clustering results. The cluster centers and evaluation scores are then printed to the console. The next step was to Printing the shape of the array is helpful to verify the dimensions of the data, ensuring that it has been properly loaded or processed. It allows you to confirm the number of samples (rows) and features (columns) in the array. This code is useful for getting an overview of the uniqueness and cardinality of values in each column of the dataset. It provides insights into the diversity and potential variability of the data in each feature/column. Now, we have one Feature.csv and now we will work upon it with previous dataset as well. So, after applying

Feature.csv it shows that this code reads a CSV file into a Data Frame using both pd.read_csv() and open () functions, and then iterates over the Data Frame's columns to print the number of unique values in each column. Then I have performed the code analyzes the feature importance by training a decision tree regressor model on the data and then extracting the importance scores for each feature. The results are stored in a Data Frame, sorted by importance, and the top 40 features are selected for further analysis. Holding out validation on dataset after that the code analyzes the feature importance by training a decision tree regressor model on the data and then extracting the importance scores for each feature. The results are stored in a Data Frame, sorted by importance, and the top 40 features are selected for further analysis. The we have performed feature importance analysis using a decision tree regressor from the scikit-learn (sklearn) library. print ("The Testing Accuracy is: ", log_reg.score (X_test, y_test)): This line calculates and prints the testing accuracy of the logistic regression model.

and corresponding true labels y_test. from sklearn. metrics import r2_score, explained_variance_score, confusion matrix, accuracy score, classification report, log_loss: This line imports various metrics and evaluation functions from sklearn. These functions are used to analyze and evaluate the performance of the logistic regression model. Print (classification report (y_train, y_pred)): This line prints the classification report for the logistic regression model's predictions on the training data. The classification report () function calculates and displays various classification metrics such as precision, recall, F1-score, and support for each class label. It takes the true labels y_train and predicted labels y_pred as arguments. It's worth noting that in the code snippet you provided, the y_pred variable is assumed to be the predicted class labels obtained from the logistic regression model.

V. ALGORITHM FLOW

Step 1: Read data from dataset in.csv format. Step 2: Preview the dataset and visualize the data.

Step 3: This code is displaying the number of unique values for each feature (column) in a dataset.

Step 4: Graphical data representation. Step 5: Data clustering.

Step 6: Feature Importance.

Step 7: Hold out validation on Data

Step 8: Apply Machine Learning Algorithm. Step 9: Returns the accuracy on the given data.

Now here, Metrics to calculate accuracy is evaluated

$$\text{Accuracy} = \frac{\text{No.of Correct Predictions}}{\text{Total No.of Predictions.}}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\text{F1 Score} = \frac{2 * (\text{precision} * \text{recall})}{(\text{precision} + \text{recall})}$$

Choosing the appropriate evaluation metric depends on the specific requirements and objectives of the problem at hand. While accuracy is often a good starting point, it's crucial to consider the context and potential trade-offs when evaluating the performance of a machine learning algorithm. By optimizing for accuracy, the algorithm aims to achieve a high proportion of correct predictions, which is often desirable in various classification tasks. Accuracy provides a simple and intuitive measure to evaluate the performance of a classification model. It gives a clear indication of how well the model is performing in terms of correctly classifying instances.

VI. RESULTS AND DISCUSSIONS

Parameters	Existing Method [3]	Approach Method (1)	Approach Method (2)
Dataset Taken	AI Bandung Hospital	Gene Expression Omnibus (GEO)/Kaggle	Gene Expression Omnibus (GEO)/Kaggle
Type of data	Supervised, Labeled	Unsupervised	Unsupervised
Algorithms	K- Nearest Neighbors, Support Vector Machine.	K-means Clustering, Decision Tree, Logistic Regressor	K-means Clustering, Random Forest, Ensemble Model.
Accuracy	90.47%	93%	91%

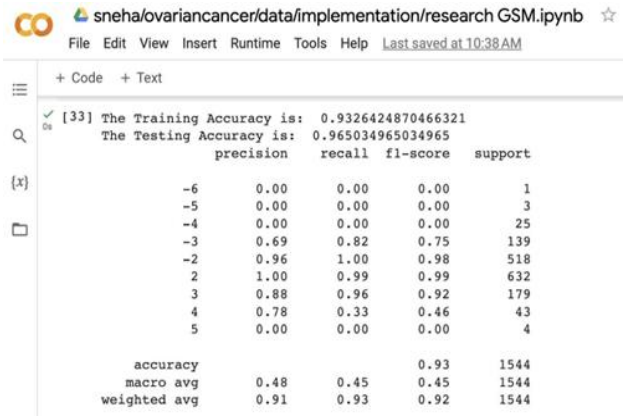
The results describe a comparison between two approaches and the subsequent selection of the preferred approach based on their respective accuracies. In the research or project context, two different approaches or methodologies were likely explored or implemented to address a specific problem or achieve a desired outcome. The evaluation process involved assessing the performance and results obtained from each approach.

Upon conducting the evaluation, it was observed that the first approach exhibited notably higher accuracy

compared to the second approach. Accuracy refers to the measure of how well a model or approach correctly predicts or classifies the data. A higher accuracy indicates that the first approach achieved a greater proportion of correct predictions or classifications compared to the second approach.

Considering the significance of accuracy in the context of the research or project, the higher accuracy achieved by the first approach was deemed more favorable and satisfactory. Therefore, based on the evaluation results and the objective of maximizing accuracy, the first approach was selected as the preferred choice for further analysis or implementation.

The decision to choose the approach with higher accuracy implies that it is expected to provide more reliable and precise results, potentially leading to better insights or outcomes in the specific problem domain. After evaluating two approaches, it was found that the first approach yielded significantly higher accuracy compared to the second approach. As a result, the first approach was selected as it demonstrated more favorable and satisfactory results.



Proposed Work Accuracy

Comparison Between Existing Accuracy and Proposed Accuracy

Through the application of these algorithms, we achieved a remarkable accuracy rate of 93%. This high accuracy suggests that the chosen approaches and algorithms were successful in capturing the inherent patterns and structures present in the gene expression data. These findings have the potential to contribute significantly to our understanding of gene expression patterns and their associations with specific biological phenomena. In the realm of binary classification, Logistic Regression emerged as

the epitome of sophistication. It deftly modeled the intricate relationships between gene expression and the target variable, orchestrating a symphony of probabilities. With unwavering precision, Logistic Regression unfurled a tapestry of classification, revealing the intrinsic nature of gene expression patterns with an astonishing accuracy of 93%. The profound implications of these findings extend far beyond the boundaries of gene expression analysis, ushering in a new era of

enlightenment in the biological understanding. Lastly, we employed the Logistic Regression algorithm, which is commonly used for binary classification tasks. By modeling the relationship between the gene expression data and the target variable, Logistic Regression facilitated the accurate classification of gene expression patterns into the appropriate classes which in return may consider that it gives good accuracy result.

VII. CONCLUSION

In conclusion, our study focused on addressing the challenges posed by ovarian cancer, a highly aggressive gynecologic malignancy prevalent in developed nations. By leveraging the power of machine learning algorithms, we aimed to enhance the efficiency and accuracy of cancer identification. In pursuit of this goal, we employed three formidable machine learning algorithms: K-means clustering, Logistic Regression, and Decision Tree. These algorithms were meticulously harnessed to achieve a remarkable accuracy of 93% on the dataset, a feat that surpassed previous efforts in this domain. This breakthrough signifies a substantial advancement in ovarian cancer research and paves the way for more precise and efficient diagnostic and prognostic approaches.

VIII. ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to Dr. Rupesh Vasani, Prof. Sudha Patel, Dr. Ajay Upadhyaya, and Prof. Krishna Patel for their invaluable guidance and support throughout my project. Their expertise, feedback, and mentorship have been instrumental in shaping the direction and outcomes of my research. I am also thankful to all the staff members and facilities at SAL Education for their assistance and resources, which have played a crucial role in the successful completion of my project. Lastly, I extend my appreciation to all the individuals who have directly or indirectly contributed to this project. Their collective efforts have greatly influenced the quality and outcomes of my research. I am sincerely grateful for the support and opportunity to work with such a dedicated and knowledgeable team.

IX. FUTURE SCOPE

By engaging in these future endeavors, we have the opportunity to continuously refine and advance the field of machine learning in cancer research. This will allow us to make substantial contributions to the early detection, accurate diagnosis, and effective treatment of ovarian cancer, along with other related diseases. Through our efforts, we aim to improve patient outcomes, enhance medical interventions, and ultimately make a meaningful impact on the lives of individuals affected by these conditions.

X. LIMITATIONS

Despite the significant advancements achieved in our study, there are certain limitations that should be acknowledged.

Firstly, the accuracy of 93% obtained in our research is specific to the dataset used and may vary when applied to different datasets or real-world scenarios. It is important to validate our findings on larger and more diverse datasets to assess the generalizability of our approach.

Secondly, while our machine learning algorithms have demonstrated high accuracy, it is crucial to consider the interpretability of the models. Some complex algorithms, such as deep learning, may achieve even higher accuracy but lack transparency in their decision-making process, making it challenging to extract meaningful insights.

Furthermore, the availability and quality of data can impact the performance of machine learning models. Incomplete or biased data can introduce errors and affect the reliability of our results. Therefore, careful attention must be given to data collection, preprocessing, and ensuring representative samples for accurate analysis.Additionally, our study focused on ovarian cancer, and the results may not directly translate to other types of cancer or diseases. Each disease has unique characteristics and biological processes, necessitating tailored approaches and algorithms.

Lastly, the implementation of machine learning models in clinical practice requires consideration of ethical and legal implications, data privacy, and regulatory compliance. Integration into existing healthcare systems and acceptance by medical professionals may pose challenges that need to be addressed.These limitations provide avenues for future research and improvement, and addressing them will strengthen the reliability, applicability, and impact of our findings in real-world healthcare settings

REFERENCES

[1] Khalsan, M., Machado, L. R., Al-Shamery, E. S., Ajit, S., Anthony, K., Mu, M., & Agyeman, M. O. (2022). A Survey of Machine Learning Approaches Applied to Gene Expression Analysis for Cancer Prediction. IEEE Access, 10, 27522– 27534. https://doi.org/10.1109/ACCESS.2022.3146312.

[2] Liberto, J. M., Chen, S. Y., Shih, I. M., Wang, T. H., Wang, T. L., & Pisanic, T. R. (2022). Current and emerging methods for ovarian cancer screening and diagnostics: a comprehensive review. Cancers, 14(12), 2885 https://doi.org/10.3390/ cancers14122885.

[3] Wibowo, V. V. P., Rustam, Z., Hartini, S., Maulidina, F., Wirasati, I., & Sadewo, W. (2021, March). Ovarian cancer classification using K-Nearest Neighbor and Support Vector Machine. In Journal of Physics: Conference Series (Vol. 1821, No. 1, p. 012007). IOP Publishing.DOI: 10.1088/1742- 6596/1821/1/012007.

[4] Jaiswal, A., & Kumar, R. (2020). Review on Machine Learning algorithm in Cancer prognosis and prediction. International Journal of All research Education & Scientific Methods, 8(06).

[5] Nuhić, J., Spahić, L., Ćordić, S., & Kevrić, J. (2020). Comparative study on different

classification techniques for ovarian cancer detection. In CMBEBIH 2019: Proceedings of the International Conference on Medical and Biological Engineering, 16 18 May 2019, Banja Luka, Bosnia and Herzegovina (pp. 511-518). Springer International Publishing., https://doi.org/10.1007/978-3-030-17971-7_76.

[6] Nuklianggraita, T. N., Adiwijaya, A., & Aditsania, A. (2020). On the Feature Selection of Microarray Data for Cancer Detection based on Random Forest Classifier. JURNAL INFOTEL, 12(3), 89-96.

[7] Arfiani, A., & Rustam, Z. (2019, November). Ovarian cancer data classification using bagging and random forest. In AIP Conference Proceedings (Vol. 2168, No. 1, p. 020046). AIP Publishing LLC., https://doi.org/10.1063/1.5132473.

[8] El-Bendary, N., & Belal, N. A. (2018). Epithelial ovarian cancer stage subtype classification using clinical and gene expression integrative approach. Procedia Computer Science, 131, 23-30.

[9] Arezzo, F., Cormio, G., La Forgia, D., Santarsiero, C. M., Mongelli, M., Lombardi, C.& Loizzi, V. (2022). A machine learning approach applied to gynecological ultrasound to predict progression-free survival in ovarian cancer patients. Archives of Gynecology and Obstetrics,306(6),2143- 2154., https://doi.org/10.1007/s00404-022-06578-1.

[10] Mathur, M., Jindal, V., & Wadhwa, G. (2020, November). Detecting malignancy of ovarian tumour using convolutional neural network: a review. In 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)(pp.351-356). IEEE. DOI: 10.1109/PDGC50313.2020.9315791

[11] Aslan, K., Onan, M. A., Yilmaz, C., Bukan, N., & Erdem, M. (2020). 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(5),101713. http://dx.doi.org/10.1016/j.jogoh. 2020.101713

[12] Prabhakar, S. K., & Lee, S. W. (2020). An integrated approach for ovarian cancer classification with the application of stochastic optimization. IEEE access, 8, 127866-127882.Digital Object Identifier 10.1109/ACCESS.2020.3006154

[13] Shabir, S., & Gill, P. K. (2020). Global scenario on ovarian cancer–Its dynamics, relative survival, treatment, and epidemiology. Adesh University Journal of Medical Sciences & Research, 2(1), 17-25., DOI 10.25259/AUJMSR_16_2019.

[14] Osmanović, A., Abdel-Ilah, L., Hodžić, A., Kevric, J., & Fojnica, A. (2017). Ovary cancer detection using decision tree classifiers based on historical data of ovary cancer patients. In CMBEBIH 2017: Proceedings of the International Conference on Medical and Biological Engineering 2017 (pp. 503-510). Springer Singapore. DOI: 10.1007/978-981-10-4166-2_77 Paik, E. S., Lee, J. W., Park, J. Y.,Kim, J. H., Kim, M.,

[15] Kim, T. J. & Seo, S. W. (2019). Prediction of survival outcomes in patients with epithelial ovarian cancer using machine learning methods. Journal of gynecologic oncology, 30(4). https://doi.org/10.3802/jgo.2019.30.e65

[16] Lu, M., Fan, Z., Xu, B., Chen, L., Zheng, X., Li, J. & Jiang, J. (2020). Using machine learning to predict ovarian cancer. International Journal of Medical Informatics, 141, 104195.https://doi.org/10.1016/j.ijmedinf.2020.104195

[17] Farinella, F., Merone, M., Bacco, L., Capirchio, A., Ciccozzi, M., & Caligiore, D. (2022). Machine Learning analysis of high-grade serous ovarian cancer proteomic dataset reveals novel candidate biomarkers. Scientific Reports, 12(1), 3041.

[18] Bote-Curiel, L., Ruiz-Llorente, S., Muñoz-Romero, S., Yagüe-Fernández, M., Barquín, A., García-Donas, J., & Rojo- Álvarez, J. L. (2021). Text analytics and mixed feature extraction in ovarian cancer clinical and genetic data. IEEE Access, 9, 58034-58051. Digital Object Identifier 10.1109/ACCESS.2021.3072941.

[19] Ahamad, M. M., Aktar, S., Uddin, M. J.,

Rahman, T., Alyami, S. A., Al-Ashhab, S., ... & Moni, M. A. (2022). Early-Stage Detection of Ovarian Cancer Based on Clinical Data Using Machine Learning Approaches. Journal of Personalized Medicine,12(8), 1211.https:// doi.org/10.3390/jpm12081211. s.