Communications in Mathematics and Applications
Vol. 16, No. 4, pp. 1093–1106, 2025
ISSN 0975-8607 (online); 0976-5905 (print)
Published by RGN Publications
DOI: 10.26713/cma.v16i4.3185

http://www.rgnpublications.com

Research Article

Feature Selection and Biomarker Identification in Ovarian Cancer Clinical Datasets Using Supervised Algorithms

Rizwan Rehman and Pinakshi Konwar*

Centre for Computer Science and Applications, Dibrugarh University, Dibrugarh, Assam, India
***Corresponding author:** pinakshikonwar@dibru.ac.in

Received: May 14, 2025 **Revised:** September 5, 2025 **Accepted:** October 29, 2025

Abstract. Ovarian Cancer is one of the most common diseases in females. It is about the unusual growth of cancer cells in the ovaries. Even with improvements in medical research and treatment, it still plays a big role in deaths caused by cancer. Early prediction and detection of this disease may save many lives. In this study, three datasets have been considered to find the features which might be useful in the prediction of this disease. In this study, an ensemble of Random Forest Classifier, XGBoost and Mutual Information Gain was applied and then voting technique was used to find the most important features. The important features from each dataset were compared with each other to get the resultant features MAF, PAX8, SERINC1, SFN, SPON1, CREBL2, ST13, and INTS5. Four machine learning techniques were used namely Random Forest, eXtreme Gradient Boosting, Light Gradient Boosting and ANN to train and evaluate using Accuracy, Precison, Recall and F1 Score. The cross-validation across independent datasets enhances the generalizability of the identified biomarkers.

Keywords. Ovarian Cancer, Machine Learning, Biomarkers

Mathematics Subject Classification (2020). 68T05, 62H30, 92C50

1. Introduction

Ovarian cancer (OC) is one of the most frequently found cancers in women. It is the second highest cause of death from common cancers that affect women, right after cervical cancer. The most common type of OC is called epithelial ovarian cancer (EOC). EOC can be divided into

three main kinds: serous OC (SOC), mucinous OC, and endometrioid OC. Among these types, SOC is the most prevalent, with high-grade serous ovarian cancer (HGSOC) making up 70% of all OC cases (Zhou *et al*. [20]). The usual treatment for EOC is surgery to remove as much tumor as possible, followed by chemotherapy that contains platinum. However, as some diseases become resistant to chemotherapy, the effectiveness of these drugs has gone down. Many new medicines are being tested in clinical trials to see how well they work for treating EOC. These include drugs that block new blood vessel growth, those that target growth factor signals, PARP inhibitors, and folate receptor inhibitors (Wang *et al*. [17]).

In recent years, artificial intelligence (AI) has become a part of medicine. AI offers a quick and accurate way to analyze large amounts of complex information and make predictions automatically. Importantly, AI systems make choices based on clinical evidence and the data provided, which tends to be more objective and efficient than how doctors usually work. The use of AI in medicine greatly lowers the chances of misdiagnosis, gives useful advice for future treatment, and eases the workload for doctors. Early detection through computational biomarker discovery could substantially improve patient survival, aligning with the goals of precision oncology. Today, more and more people are learning about computer-aided methods, and applying this knowledge in everyday medical practice shows good results (Zhou *et al*. [20]). AI methods are being used more and more these days for accurate diagnosis of many diseases. Recently, different AI tools, especially machine learning (ML) and deep learning, have become well-liked for diagnosing and predicting many diseases, especially cancer, because they have great benefits. However, few studies have looked at how AI (ML) tools can predict ovarian cancer. Because of the limits of these studies, there is a clear need for new and better research (Ayyoubzadeh *et al*. [4]).

Machine learning, a subset of artificial intelligence, has gained prominence in medical informatics due to its ability to process large datasets and uncover patterns that are not easily discernible through traditional statistical methods. In the context of ovarian cancer, machine learning algorithms can be employed to analyze complex medical data, such as genetic information, imaging results, and patient histories, to predict cancer presence more accurately and earlier than conventional methods (Sundari and Brintha [15]).

In this research, an Ensemble of machine learning techniques, namely Random Forest Classifier, XGBoost and Mutual Information Gain have been used for feature extraction.

2. Related Work

According to Ayyoubzadeh *et al*. [4], artificial intelligence tools to predict ovarian cancer by analyzing blood test results and tumor markers, identifying the random forest model as the most accurate with over 86% accuracy, and suggesting that AI can aid in early and cost-effective diagnosis.

According to Hira *et al*. [7], a systematic review highlights the growing application of deep learning (DL) in ovarian cancer (OC) diagnosis and analysis. While most studies focus on detection and diagnosis using imaging data still, there is a lack of research on prediction and prevention.

Hamidi *et al*. [5] suggests AI-based approaches have demonstrated the ability to predict clinical stage, histotype, and residual tumor burden preoperatively. The integration of miRNA biomarkers with machine learning models has yielded promising results, with some studies reporting area under the ROC curve values of up to 100% for ovarian cancer prediction.

Ahamad *et al*. [2] apply machine learning models along with statistical methods to clinical data obtained from 349 patients individuals to conduct predictive analytics for early diagnosis of ovarian cancer.

Juwono *et al*. [8] propose an optimal simultaneous feature weighting and parameter optimization approach using adaptive differential evolution (ADE) with LASSO regularization to detect ovarian cancer with high accuracy.

Liu *et al*. [9] identifies and validates nine diagnostic characteristic genes for ovarian cancer using bioinformatics and machine learning.

Hamidi *et al*. [6] identifies 10 miRNAs as potential biomarkers for ovarian cancer and demonstrates their high accuracy in diagnosing the disease using machine learning models, suggesting that the serum miRNA profile is a promising diagnostic tool.

Anaissi *et al*. [3] demonstrates ensemble feature learning using support vector machines (SVMs) has shown promise in genomic data classification. ESVM-RFE, an ensemble approach combining SVM with recursive feature elimination, outperformed traditional methods in gene selection and classification of microarray data.

Table 1. Comparative summary of prior studies on ovarian cancer prediction

Study (Year)	Dataset / Sample type	Algorithms used	Reported accuracy / AUC	Key contribution or novelty
Ahamad *et al*. [2]	Clinical dataset with 349 patient samples	Random Forest, GBM, LGBM	96% accuracy	Demonstrated clinical data-based ML prediction for early-stage ovarian cancer diagnosis
Hamidi *et al*. [6]	Serum miRNA dataset	SVM, Random Forest, XGBoost	AUC ≈ 0.97-1.00	Identified 10 circulating miRNAs as promising biomarkers for early OC detection
Hamidi *et al*. [5]	Circulating miRNA profiles	Boruta, XGBoost	Accuracy ≈ 93-95%	Applied Boruta-based feature selection to identify miRNA biomarkers for diagnosis and prognosis
Juwono *et al*. [8]	Genomic expression dataset	Adaptive Differential Evolution (ADE) with LASSO regularization	95% accuracy	Proposed optimized feature weighting with ADE-LASSO hybrid for OC classification
Hira *et al*. [7]	Multiple OC imaging and gene expression datasets (systematic review)	Deep Learning (CNN, Autoencoder)	—	Provided a systematic synthesis of DL-based OC prediction, identified research gaps in non-imaging genomic ML

Table 1 (*continued*)

Study (Year)	Dataset / Sample type	Algorithms used	Reported accuracy / AUC	Key contribution or novelty
Sundari and Brintha [15]	Clinical and imaging data	SVM, Random Forest, XGBoost	98.7% (XGBoost)	Compared traditional and ensemble ML methods for tumor classification
Zhou *et al*. [20]	TCGA mRNA and proteomic data	Deep Neural Networks	95% (F1-score)	Applied weakly supervised learning for biomarker-guided treatment prediction

3. Methodology

3.1 Datasets

This research consists of three sets of data. The first dataset is the High-Grade Serous Ovarian Cancer (TCGA, GDC) obtained from cBioPortal. It consists of around 489 patient tumor samples. The mRNA Expression (RNA-seq) dataset was considered from it. The experimental code supporting this study is accessible at https://github.com/pinakshikonwar/Feature-Selection-using-ensemble-ML-Techniques.

The second dataset is the Filtering_pyComBat dataset from Kaggle. The dataset consists of about 594 genes and 4181 patient samples.

The third dataset comes from NCBI. This project seems to learn about how tumors change within the same patient by studying different sampling areas (both primary and metastatic) at the time of diagnosis and when the disease comes back. For this, 183 biopsies from 50 patients were collected, and bulk mRNA sequencing was done.

3.2 Data Analysis

The following steps were taken during data analysis which are shown in Figure 1 and Figure 2 which can be explained as:

3.2.1 Data Preprocessing and Labelling

Getting data ready for applications in machine learning includes two important steps: preprocessing and labeling. Preprocessing means cleaning the data by fixing missing information, scaling the features, and putting the data in a consistent format. Labeling involves a trained model being applied to another model to predict the target variables.

Here ANN was applied to train and label similar datasets.

3.2.2 Balancing using SMOTE

The datasets were balanced using SMOTE(Synthetic Minority Over-sampling Technique) to enhance the results. Unbalanced datasets are biased and overfitting might occur. So, it is necessary to balance the dataset before training to give good results.

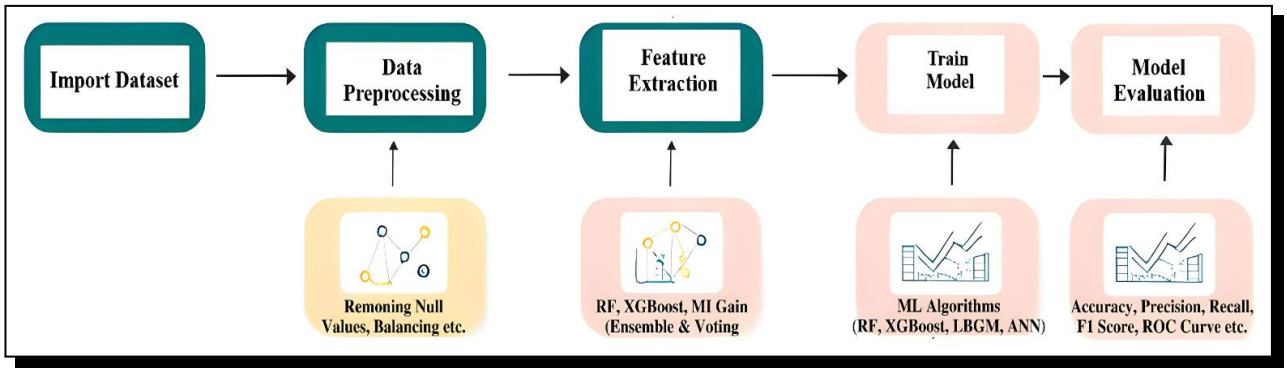

Figure 1. Methodology

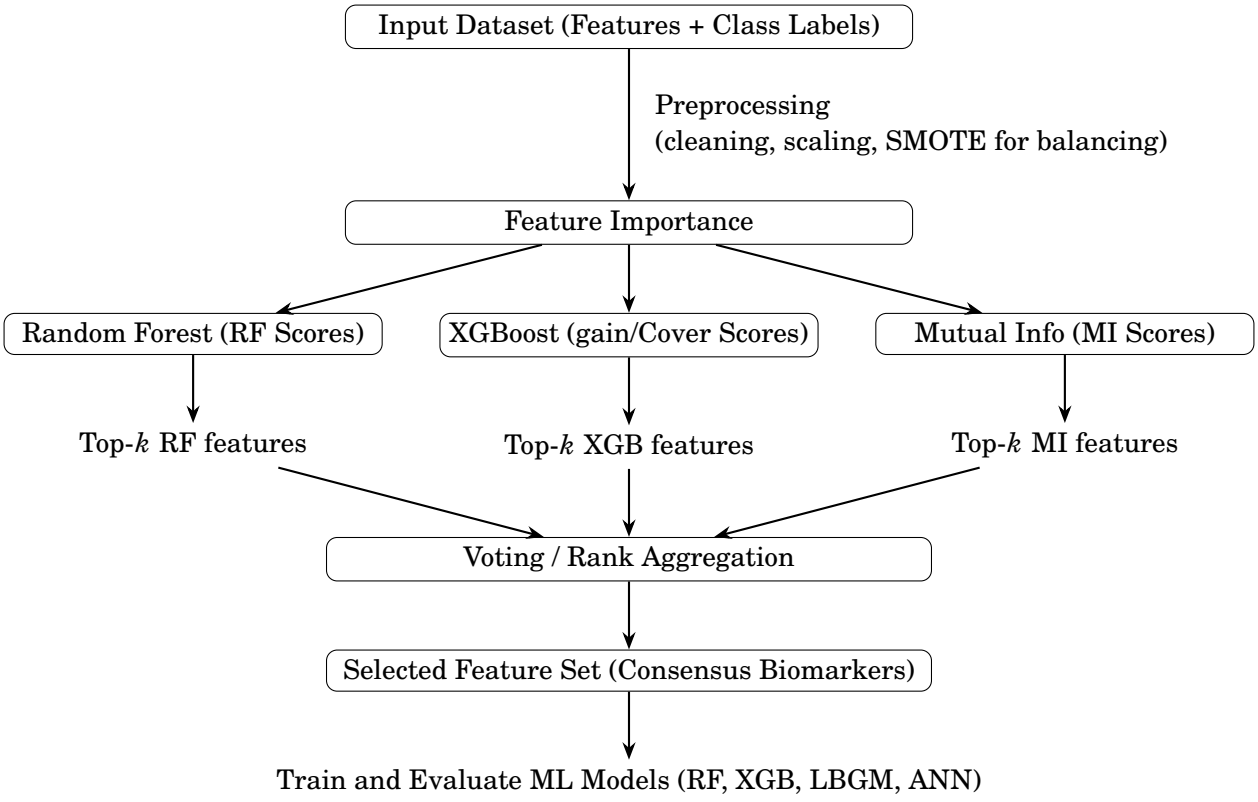

Figure 2. Flow diagram of methodology

3.2.3 Feature Extraction

Feature extraction is the process of selecting the best features from a large number of features. In this study we have found out the best features from the three datasets and then compared with each other and finally eight features were selected to be the best. To select the features we have used an ensemble of Random Forest Classifier (RFC), eXtreme Gradient Boosting (XGBoost) and Mutual Information Gain (MI) methods and then used a voting mechanism to select the best features from each dataset.

Mandal *et al*. [11])uses a three-stage wrapper-filter method for selecting important features in medical reports to detect diseases. It is better than the best current methods because it finds relevant features and lowers the number of medical tests and costs.

Random Forest Classifier

A Random Forest considers a feature important if using it to split the data helps to group the data (by the target classes) more effectively than other features.

Recent studies have looked into ways to choose and extract features to better classify cancer with machine learning. Random Forest (RF) has shown good results in diagnosing breast cancer. Researchers have created a better method using RF to extract rules for clearer classification (Wang *et al*. [18]).

XGBoost

XGBoost is a well-known method for boosting that helps make predictions better. It gives us a way to see which features are important, often using a measure called gain. When XGBoost adds a split to one of its trees, it figures out how much that split helps the model perform better (Shehzadi *et al*. [13]) develops machine learning models using RNA sequences to extract cancer-related biological information, with the XGBoost classifier showing high accuracy and sensitivity, useful for precision medicine, drug discovery, and clinical oncology.

Mutual Information Gain

Mutual information looks at how a feature relates to a target without being tied to any specific model. This idea comes from information theory. It shows how much knowing one variable, like a feature, helps you understand another variable, which is the target. In simple terms, mutual information looks at how knowing the feature can reduce uncertainty about the target.

A new method that uses mutual information, LASSO, and a genetic algorithm is suggested for identifying breast cancer. This method reaches a 96% accuracy in classification using just 23 features from the van't Veer dataset, which is better than the best current models (Abd-elnaby *et al*. [1]).

3.2.4 Training Machine Learning Algorithms

The model is trained using various machine learning algorithms such as Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Method (LGBM) and Artificial Neural Networks (ANN) (Mohapatra *et al*. [12]) explains how Machine Learning can help predict breast cancer in women. It compares how well different methods work, including Support Vector Classifier, Random Forest, and XGBoost. XGBoost performed the best with an accuracy of 98.7%.

Ahamad *et al*. [2] uses machine learning and statistics on clinical data from 349 patients to find important markers and accurately tell apart benign from malignant ovarian cancer using models like RF, GBM, and LGBM.

3.2.5 Model Evaluation

Evaluation metrics are very important to figuring out how well machine learning models work. These metrics give a clear way to see how good a model is, helping researchers and users choose and improve models wisely. In this study, the most popular evaluation metrics: accuracy, precision, recall, F1 score, and the ROC curve are used.

4. Experimental Results

Using an Ensemble of Random Forest, XGBoost, and Mutual Information Gain, we found a group of important features from the CBio dataset as shown in Figure 3. MAF was the most important feature, having a much higher importance score than the others. Followed byRPS12, GTPBP3, AKT3, RPS3A, RPL6, NSA2, SERINC1, RNF13, PAX8, CAV1, CREBL2, CSDE1, ST13, ARPC5L, QKI, DAB2, ITM2B, SPON1, COPS4, SFN, TIMP2, SLC12A7, YEATS2, HDGF, UBL3, FAS, CELF2, SLC35C2, and BNC2. These genes might be considered important biomarkers for ovarian cancer.

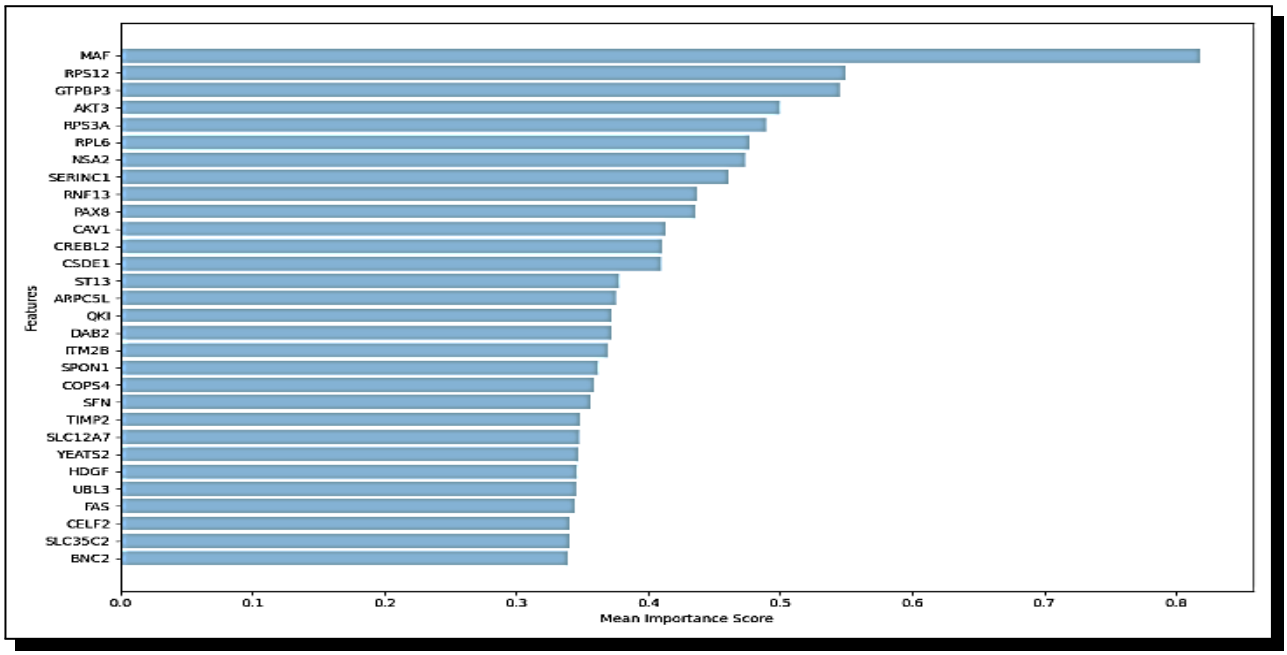

Figure 3. Important features using the CBio dataset

The results from the feature selection show that all the machine learning models performed well, with the Artificial Neural Network (ANN) the best. ANN reached an accuracy of 0.97, with precision and recall both at 0.98, and an F1 score of 0.97. This means it has strong predictive ability and a good balance across different measures.Traditional models like Random Forest (RF) and LightGBM (LGBM) also showed good performance, each achieving 0.94 accuracy, while XGBoost was close behind with 0.93 accuracy as given in Table 2. These findings indicate that the chosen features are very useful and help improve classification results across various model types.

Table 2. Evaluation of results using machine learning techniques after feature selection using CBio dataset

Methods	Accuracy	Precision	Recall	F1 score
RF	0.94	0.94	0.93	0.92
XGBoost	0.93	0.93	0.93	0.92
LGBM	0.94	0.94	0.94	0.93
ANN	0.97	0.98	0.98	0.97

Figure 4 shows a bar plot of the top features identified through an ensemble method combining a Random Forest Classifier (RFC), XGBoost, and Mutual Information Gain using the Kaggle Dataset.MAF is at the top again, showing it is very relevant both the datasets. Other important genes are PPOX, FAM153B, INTS5, MT1F, S100A13, ST13, VEGFA, EIF1, VPS52, RPS25, DYRK2, MEF2C, FAM136A, GTF2B, SFN, IL6ST, PPP2CA, FEZ2, GCOM1, FEN1, SPON1, CLIP4, CCDC86, PAX8, SERINC1, CREBL2, CD47, CRK, and OAS2, which have biological importance.

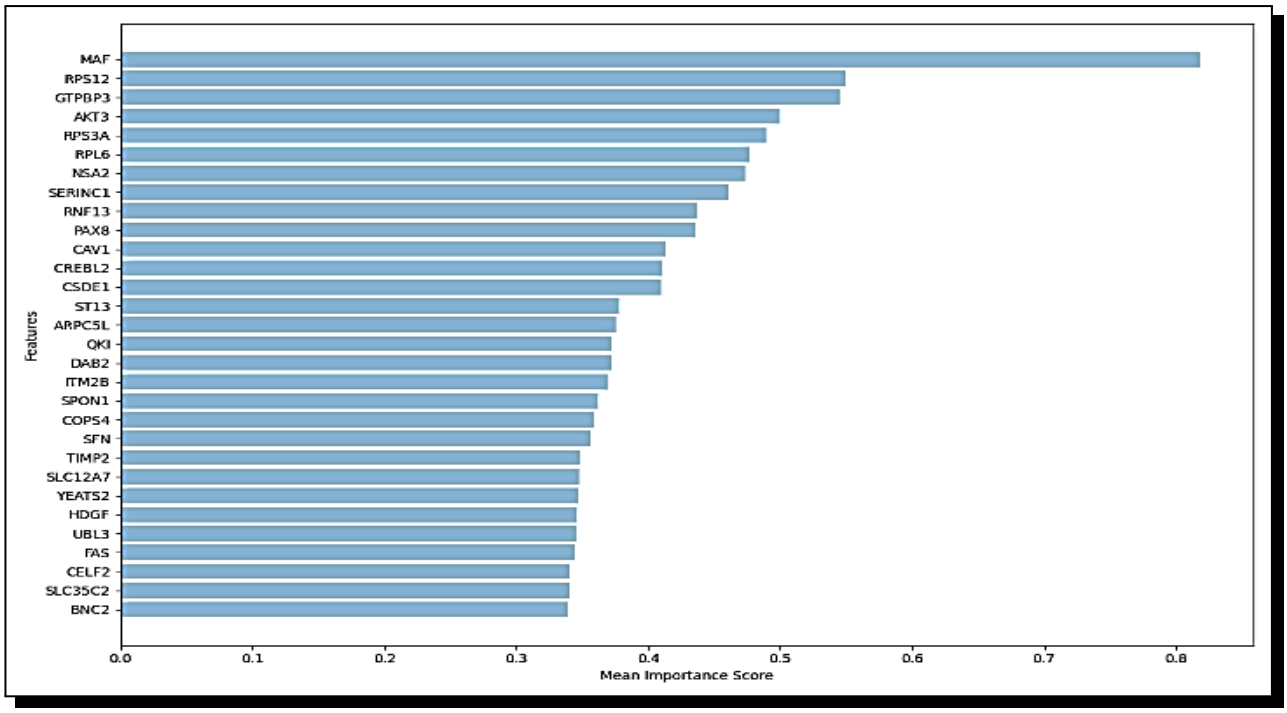

Figure 4. Important features using the Kaggle dataset

Table 3. Evaluation of results using machine learning techniques after Feature Selection using Kaggle Dataset

Methods	Accuracy	Precision	Recall	F1 score
RF	0.96	0.97	0.97	0.96
XGBoost	0.98	0.98	0.98	0.98
LGBM	0.97	0.97	0.98	0.97
ANN	0.97	0.98	0.97	0.98

The results from the feature selection given in Table 3 shows that all four machine learning models performed well in classifying data. XGBoost was the best, scoring perfectly on all measures, 0.98 for accuracy, precision, recall, and F1 score, highlighting its ability to identify patterns from the chosen features.

The Artificial Neural Network (ANN) and LightGBM (LGBM) also performed very well, each getting 0.97 accuracy with a good balance between precision and recall. Random Forest (RF) was just behind with an accuracy of 0.96 and strong prediction results.

The bar graph in Figure 5, shows the main features found using an ensemble of methods that include Random Forest Classifier (RFC), XGBoost, and Mutual Information Gain using the GEO dataset. This method helps in selecting features in a more complete and fair way by using the best parts of each technique.

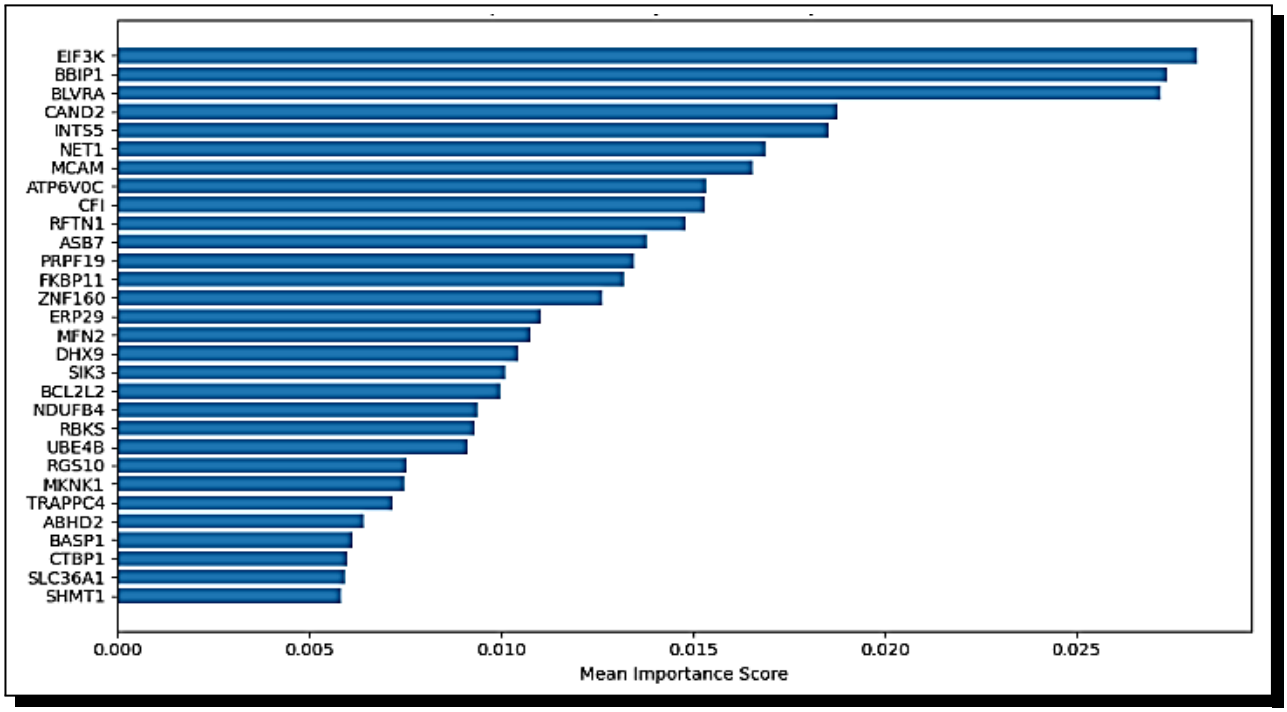

Figure 5. Important features using the GEO dataset

Table 4. Evaluation of results using machine learning techniques after feature selection using GEO dataset

Methods	Accuracy	Precision	Recall	F1 score
RF	0.56	0.61	0.57	0.57
XGBoost	0.46	0.50	0.47	0.46
LGBM	0.51	0.57	0.51	0.52
ANN	0.52	0.57	0.51	0.52

The assessment of machine learning models following feature selection shows fairly average results for all methods. The Random Forest (RF) model performed the best, reaching an accuracy of 0.56 and a slightly improved precision of 0.61 given in Table 4.

A heatmap of the common features in between the CBio and Kaggle Datasets is given in Figure 6. The common features are CREBL2, MAF, PAX8, SERINC1, SFN, SPON1, ST13.

Similarly, comparing the resultant features between the Kaggle and GEO datasets gives a feature INTS5 whose importance score is given in Figure 7.

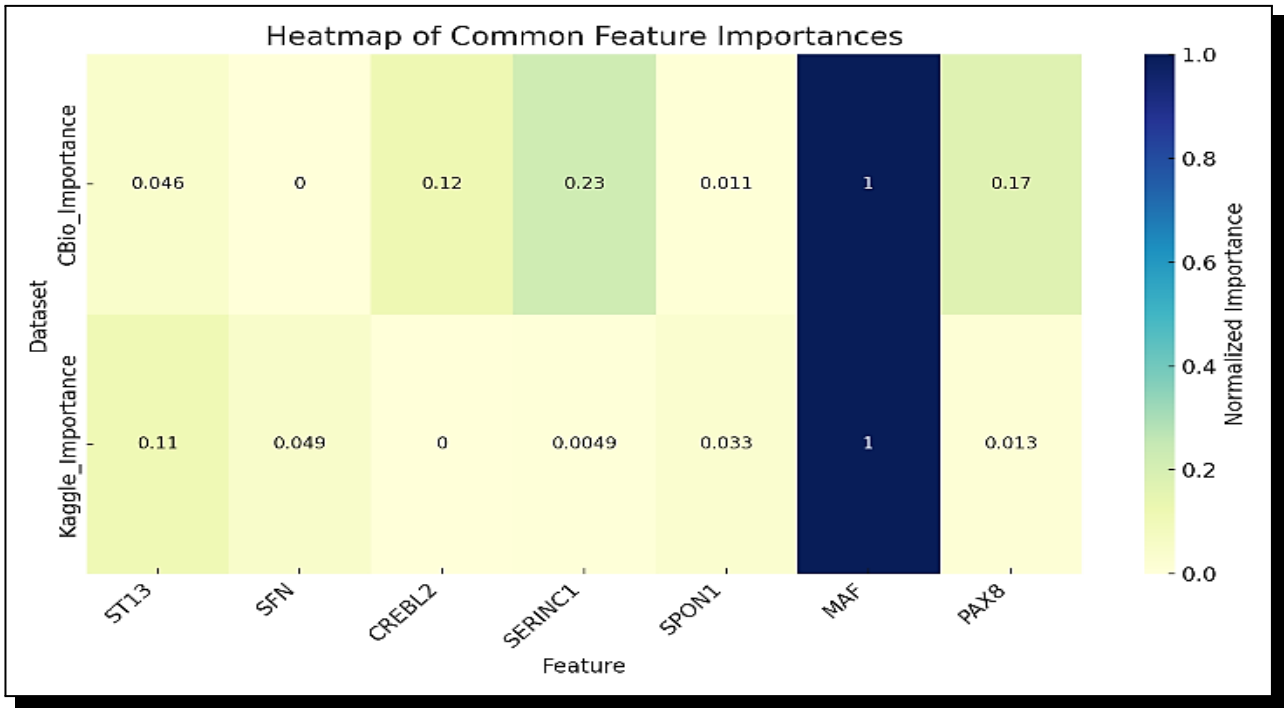

Figure 6. Heatmap of the common features between CBio and Kaggle dataset

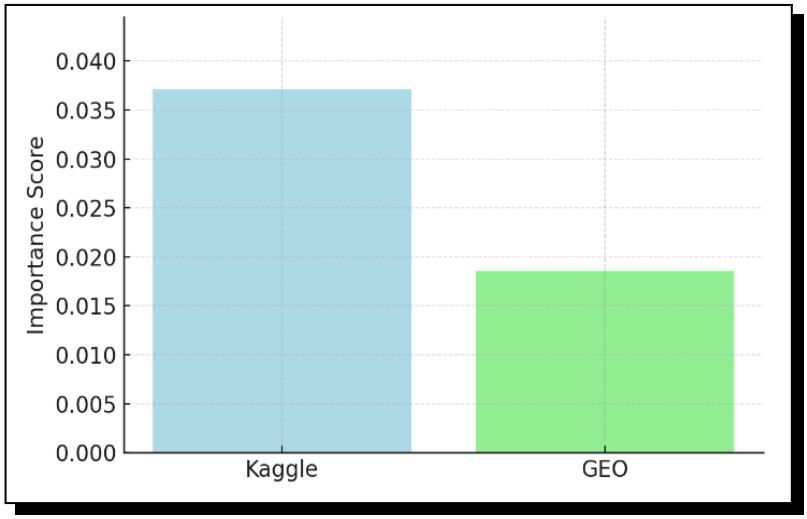

Figure 7. Importance score of INTS5 across the two datasets

5. Discussion

From the above experimental results, it is clear that MAF, PAX8, SERINC1, SFN, SPON1, CREBL2, ST13, and INTS5 are important features in predicting Ovarian Cancer.

MAF (c-Maf Transcription Factor) is a basic leucine zipper transcription factor. In ovarian cancer, when c-Maf is too much, it is linked to more serious illness. This may happen because it helps the cancer cells grow, spread, and create a harmful environment for the immune system. High levels of c-Maf are connected to bad results, suggesting it might help the cancer grow. A high expression of tumor MAF is linked to a shorter overall survival for patients with ovarian cancer. This means c-Maf is a prognostic biomarker to predict severe disease (Liu *et al*. [10]).

PAX8 (Paired Box 8) has a critical role in ovarian cancer cell survival and metastasis. It acts as a key regulatory transcription factor in these tumors (Soriano *et al*. [14]). It is a well-established diagnostic biomarker in pathology. Its presence, detected by immunohistochemistry, is used to confirm if a carcinoma is of ovarian or fallopian or endometrial origin (Wang *et al*. [19]).

SERINC1 encodes a membrane protein involved in lipid synthesis. It is expressed at low to moderate levels in most normal tissues. In ovarian cancer, SERINC1 expression is often elevated. Analyses of patient datasets indicate that high SERINC1 mRNA in ovarian tumors is associated with significantly shorter survival, implying that aggressive tumors tend to have SERINC1 upregulation (The Human Protein Atlas [16]).

SFN (Stratifin / 14-3-3σ) helps cancer cells survive and resist chemotherapy. Tumors with high SFN tend to be more resilient and harder to treat. Because of this, SFN may act as both a prognostic and predictive biomarker, giving doctors clues about how a patient might respond to standard treatments and whether alternative approaches might be needed.

SPON1 builds part of the structure around cancer cells, helping them grow and spread. It's often found at high levels in ovarian tumors, and patients with elevated SPON1 tend to have earlier relapses. That makes it a promising biomarker for prognosis, helping doctors spot cases that might benefit from more intensive therapy or new treatment options.

CREBL2 tends to go quiet in ovarian cancer, often due to gene loss. While it's not widely used in testing yet, its consistent disappearance from tumor cells suggests it could serve as a warning sign of more advanced or aggressive cancer. Scientists are exploring whether its absence could help flag cancers early or identify patients who may need closer monitoring

ST13 normally helps control how cells grow and handle stress. When tumors lose ST13, they may become more aggressive and less responsive to chemo. So, measuring ST13 levels could help predict how a patient's cancer might behave, and low levels might signal the need for novel or experimental treatments.

INTS5 is a gene that plays a role in how cells process RNA, and recent studies suggest it could be important in ovarian cancer. Research has shown that INTS5 tends to be more active in ovarian tumor tissues compared to normal ones, which hints at its potential as a diagnostic marker. More importantly, higher levels of INTS5 have been linked to poorer survival outcomes in patients, meaning it could help predict how aggressive the disease might be.

The study identified eight genes MAF, PAX8, SERINC1, SFN, SPON1, CREBL2, ST13, and INTS5, that appeared consistently across the analyzed datasets. Each gene is involved in pathways that regulate cell growth, differentiation, and stress response. MAF and PAX8 influence transcription and tumour cell proliferation, while SERINC1 and SFN contribute to metabolic activity and resistance to chemotherapy. SPON1 supports cell adhesion and metastatic behaviour, whereas CREBL2, ST13, and INTS5 participate in protein folding and RNA processing linked to tumour progression. Together, these genes represent biologically significant markers that may assist in early detection and therapeutic assessment of ovarian cancer.

6. Conclusion and Future Work

This paper focuses on obtaining potential biomarkers for ovarian cancer using three datasets. After a series of experiments the suitable features from these datasets were found to be MAF, PAX8, SERINC1, SFN, SPON1, CREBL2, ST13, and INTS5. These genes are functionally associated with tumour development, stress response, and metastasis, suggesting their potential utility in early diagnosis and therapeutic decision-making. The proposed approach demonstrates that integrating multiple algorithms and datasets enhances both model accuracy and the biological interpretability of results. Future work will focus on biological validation of these biomarkers through experimental studies and the inclusion of multi-omics data such as proteomic and methylation profiles to strengthen clinical translation. Further improvements may include dataset augmentation using larger and more heterogeneous cohorts to enhance model robustness and generalizability. Additionally, the application of transfer learning can extend this framework to other related cancers, facilitating cross-cancer comparison and discovery of shared molecular signatures. Incorporating explainable AI methods will also aid in translating computational findings into clinically interpretable insights, contributing toward precision oncology and data-driven treatment planning.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

All the authors contributed significantly in writing this article. The authors read and approved the final manuscript.

References

[1] M. Abd-elnaby, M. Alfonse and M. Roushdy, A hybrid mutual information-lasso-genetic algorithm selection approach for classifying breast cancer, in: *Digital Transformation Technology*, D. A. Magdi, Y. K. Helmy, M. Mamdouh and A. Joshi (editors), Lecture Notes in Networks and Systems, Volume 224, Springer, Singapore (2022), DOI: 10.1007/978-981-16-2275-5_36.

[2] M. M. Ahamad, S. Aktar, M. J. Uddin, T. Rahman, S. A. Alyami, S. Al-Ashhab, H. F. Akhdar, A. K. M. Azad and M. A. Moni, Early-stage detection of ovarian cancer based on clinical data using machine learning approaches, *Journal of Personalized Medicine* **12**(8) (2022), 1211, DOI: 10.3390/jpm12081211.

[3] A. Anaissi, M. Goyal, D. R. Catchpoole, A. Braytee and P. J. Kennedy, Ensemble feature learning of genomic data using support vector machine, *PLoS ONE* **11**(6) (2016), e0157330, DOI: 10.1371/journal.pone.0157330.

[4] S. M. Ayyoubzadeh, M. Ahmadi, A. B. Yazdipour, F. Ghorbani-Bidkorpeh and M. Ahmadi, Prediction of ovarian cancer using artificial intelligence tools, *Health Science Reports* **7**(7) (2024), e2203, DOI: 10.1002/hsr2.2203.

[5] F. Hamidi, N. Gilani, R. A. Belaghi, H. Yaghoobi, E. Babaei, P. Sarbakhsh and J. Malakouti, Identifying potential circulating miRNA biomarkers for the diagnosis and prediction of ovarian cancer using machine-learning approach: Application of Boruta, *Frontiers in Digital Health* **5** (2023), 1187578, DOI: 10.3389/fdgth.2023.1187578.

[6] F. Hamidi, N. Gilani, R. A. Belaghi, P. Sarbakhsh, T. Edgünlü and P. Santaguida, Exploration of potential miRNA biomarkers and prediction for ovarian cancer using artificial intelligence, *Frontiers in Genetics* **12** (2021), Article 724785, DOI: 10.3389/fgene.2021.724785.

[7] M. T. Hira, M. A. Razzaque and M. Sarker, Ovarian cancer data analysis using deep learning: A systematic review, *Engineering Applications of Artificial Intelligence* **138** Part A (2024), 109250, DOI: 10.1016/j.engappai.2024.109250.

[8] F. H. Juwono, W. K. Wong, H. T. Pek, S. Sivakumar and D. D. Acula, Ovarian cancer detection using optimized machine learning models with adaptive differential evolution, *Biomedical Signal Processing and Control* **77** (2022), 103785, DOI: 10.1016/j.bspc.2022.103785.

[9] J. Liu, L. Liu, P. A. Antwi, Y. Luo and F. Liang, Identification and validation of the diagnostic characteristic genes of ovarian cancer by bioinformatics and machine learning, *Frontiers in Genetics* **13** (2022), Article 858466, DOI: 10.3389/fgene.2022.858466.

[10] M. Liu, Z. Tong, C. Ding, F. Luo, S. Wu, C. Wu, S. Albeituni, L. He, X. Hu, D. Tieri, E. C. Rouchka, M. Hamada, S. Takahashi, A. A. Gibb, G. Kloecker, H. G. Zhang, M. Bousamra, B. G. Hill, X. Zhang and J. Yan, Transcription factor c-Maf is a checkpoint that programs macrophages in lung cancer, *The Journal of Clinical Investigation* **130**(4) (2020), 2081 – 2096, DOI: 10.1172/JCI131335.

[11] M. Mandal, P. K. Singh, M. F. Ijaz, J. Shafi and R. Sarkar, A tri-stage wrapper-filter feature selection framework for disease classification, *Sensors* **21**(16) (2021), 5571, DOI: 10.3390/s21165571.

[12] S. K. Mohapatra, A. Jain, Anshika and P. Sahu, Comparative approaches by using machine learning algorithms in breast cancer prediction, in: *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering* (ICACITE, Greater Noida, India, 2022), pp. 1874 – 1878, (2022) DOI: 10.1109/ICACITE53722.2022.9823470.

[13] T. Shehzadi, A. Majid, M. Hameed, A. Farooq and A. Yousaf, Intelligent predictor using cancer-related biologically information extraction from cancer transcriptomes, in: *2020 International Symposium on Recent Advances in Electrical Engineering & Computer Sciences* (RAEE & CS, Islamabad, Pakistan, 2020), pp. 1 – 5, (2020), DOI: 10.1109/RAEECS50817.2020.9265692.

[14] A. A. Soriano, T. de Cristofaro, T. Di Palma, S. Dotolo, P. Gokulnath, A. Izzo, G. Calì, A. Facchiano and M. Zannini, PAX8 expression in high-grade serous ovarian cancer positively regulates attachment to ECM via Integrin $\beta 3$, *Cancer Cell International* **19** (2019), Article number: 303, DOI: 10.1186/s12935-019-1022-8.

[15] M. J. Sundari and N. C. Brintha, A comparative study of various machine learning methods on ovarian tumor, in: *2021 Sixth International Conference on Image Information Processing* (ICIIP, Shimla, India, 2021), pp. 314 – 319, (2021), DOI: 10.1109/ICIIP53038.2021.9702697.

[16] The Human Protein Atlas (2022), *SERINC1 in Ovarian Cancer*, online accessed on June 6, 2025, URL: https://www.proteinatlas.org/ENSG00000111897-SERINC1.

[17] C.-W. Wang, Y.-C. Lee, C.-C. Chang, Y.-J. Lin, Y.-A. Liou, P.-C. Hsu, C.-C. Chang, A.-K.-O. Sai, C.-H. Wang and T.-K. Chao, A weakly supervised deep learning method for guiding ovarian cancer treatment and identifying an effective biomarker, *Cancers* **14**(7) (2022), 1651, DOI: 10.3390/cancers14071651.

[18] S. Wang, Y. Wang, D. Wang, Y. Yin, Y. Wang and Y. Jin, An improved random forest-based rule extraction method for breast cancer diagnosis, *Applied Soft Computing* **86** (2020), 105941, DOI: 10.1016/j.asoc.2019.105941.

[19] Y. Wang, Y. Wang, J. Li, Z. Yuan, B. Yuan, T. Zhang, J. M. Cragun, B. Kong and W. Zheng, PAX8: A sensitive and specific marker to identify cancer cells of ovarian origin for patients prior to neoadjuvant chemotherapy, *Journal of Hematology & Oncology* **6** (2013), Article number: 60, DOI: 10.1186/1756-8722-6-60.

[20] J. Zhou, W. Cao, L. Wang, Z. Pan and Y. Fu, Application of artificial intelligence in the diagnosis and prognostic prediction of ovarian cancer, *Computers in Biology and Medicine* **146** (2022), 105608, DOI: 10.1016/j.compbiomed.2022.105608.