RESEARCH ARTICLE

Editorial Process: Submission:10/12/2024 Acceptance:03/15/2025

Enhancing Personalized Chemotherapy for Ovarian Cancer: Integrating Gene Expression Data with Machine Learning

Mahmood Khalsan[1]*, Fawaz Al-Alloosh[1], Ahmed SK Al-Khafaji[1,2]

Abstract

Objective: Ovarian cancer's complexity and heterogeneity pose significant challenges in treatment, often resulting in suboptimal chemotherapy outcomes. This study aimed to leverage machine learning algorithms, gene selection, and gene expression data to improve chemotherapy results. **Methods:** The mutual_info_classif approach was employed to identify the most informative genes for predicting treatment responses. Ten machine learning techniques were used to assess and optimize the predictive potential of these genes. **Result:** By examining the reciprocal relationships between gene expression and chemotherapy outcomes, the study identified a subset of 20 critical genes essential for treatment efficacy. Among the selected genes, the Random Forest classifier demonstrated the highest accuracy, achieving 97% accuracy, 98% precision, 97% recall, and a 97.5% F1-score in predicting treatment responses. With statistical significance ($p = 0.019$), the carboplatin predictor successfully distinguished between platinum-sensitive and platinum-resistant patients. Additionally, the combined predictor for the platinum-taxane regimen revealed a significant difference in survival between predicted responders and non-responders, with median survival times of 12.9 months and 8.1 months, respectively ($p < 0.045$). **Conclusion:** The exceptional performance of this model highlights its ability to integrate complex gene expression data, facilitating the development of personalized chemotherapy regimens.

Keywords: Personalized Chemotherapy- Gene Expression Data- Biomarkers- Machine Learning

Asian Pac J Cancer Prev, **26 (3)**, 959-967

Introduction

Ovarian cancer is one of the most common cancers worldwide [1], known for its poor prognosis and high mortality rates [2]. Due to late-stage detection and inconsistent treatment outcomes, ovarian cancer remains a challenging illness to treat, with a high fatality rate. Although advancements in chemotherapy and surgery have provided some improvements, there is still a pressing need for personalized treatment options. In this context, novel machine learning (ML) algorithms have been developed to enhance tailored treatment plans, fueled by recent advancements in computational techniques and genomics. Specifically, integrating ML with gene expression data has become a cutting-edge approach to better individualize treatment for each patient.

Analyzing gene expression data offers new insights into the molecular makeup of ovarian cancer. By studying the activity of hundreds of genes, researchers can identify new patterns associated with drug sensitivity and resistance. These patterns contribute to the creation of personalized medicine regimens [3], providing a path to more targeted and effective treatments by illuminating key processes involved in drug response and resistance [4].

ML approaches are increasingly used to analyze complex gene expression data, which is characterized by high dimensionality. Advanced ML algorithms, such as support vector machines (SVM), random forests (RF), and neural networks, can process these data more accurately, predicting how patients will respond to specific drugs [5]. Recently, ML algorithms have effectively integrated omics data to develop predictive models, which could revolutionize chemotherapy selection and delivery [6].

Khalid et al. [7] demonstrated how ML approaches can utilize gene expression data to predict chemotherapy response, leading to more personalized treatment plans. Similarly, Jiang et al. [8] showed that combining ML with multi-omics data could improve therapy efficacy and prediction accuracy.

Personalized chemotherapy is an intriguing treatment option for ovarian cancer, known for its heterogeneity and variable response to therapies. Traditional methods of individualized chemotherapy have struggled to capture the molecular complexity of patients, as they are based on clinical and histopathological criteria. Recently, the focus has shifted toward integrating molecular data to enhance therapeutic approaches.

Yu et al. [9] employed various machine learning

*[1]Scientific Department, Warith International Cancer Institute, Karbala, 56001, Iraq. [2]Department of Biology, College of Science, University of Baghdad, Al-Jadriya, Baghdad 10071, Iraq. *For Correspondence: mahmoud@mizan.eduiq*

techniques to predict platinum drug responses. These models were evaluated using leave-one-out cross-validation and included data from 130 ovarian serous carcinoma patients. The random forest classifier achieved the highest accuracy (79%) compared to other classifiers in the study.

Fekete et al. [10] identified predictive biomarkers for chemotherapy resistance in ovarian cancer. The study focused on the platinum and taxane regimen and utilized GEO and TCGA data to compare treatment responders and non-responders. The researchers identified eight significant genes linked to resistance: AKIP1, MARVELD1, AKIRIN2, CFL1, SERBP1, PDXK, TFE3, and NCOR2. Additionally, a combined dataset was created to discover new biomarkers. Similarly, Buttarelli et al. [11] developed a method to identify key biomarkers associated with treatment outcomes by employing gene expression profiles to predict responses to chemotherapy in ovarian cancer. Their study identified four downregulated genes and six upregulated genes, which were used as identifiers for training a Random Forest (RF) approach. The RF model achieved 93% accuracy and 94% precision.

Weberpals et al. [12] conducted an extensive analysis that categorized ovarian cancer into molecular subtypes based on gene expression patterns. Their study provided insights into the varied responses to chemotherapy, aiming to use molecular signatures to guide treatment selection. In their examination of 39 patient samples, they found that platinum-resistant (PR) tumors displayed distinct mutations and amplifications, whereas platinum-sensitive (GR) tumors were associated with longer median progression-free intervals and BRCA2 mutations. Additionally, gene expression analysis of GR samples revealed higher levels of immune cell infiltration and PD-L1 expression compared to PR samples.

Huan et al. [13] developed a machine learning-derived prognostic signature (MLDPS) for clinical assessments in ovarian cancer, highlighting the limited effectiveness of current models in predicting outcomes and the need for better biomarkers. The study identified robust prognostic risk genes that outperformed existing models and suggested that patients with low MLDIS scores might benefit from immunotherapy and chemotherapy. This tool could significantly improve precision treatment and clinical management for ovarian cancer patients.

Lu et al. [14] developed an ML model for predicting relapse following first-line chemotherapy. They created a pharmacological model using the GSE9891 dataset from TCGA and validated it using data from the Cancer Cell Line Encyclopedia (CCLE). Their 10-gene predictive model indicated that patients who responded well to treatment had higher recurrence-free survival times, suggesting that those who responded partially or poorly would benefit from switching to other medications.

Hsu et al. [15] examined five machine learning models for predicting muscle loss, analyzing changes in the skeletal muscle index (SMI) in 617 ovarian cancer patients who underwent chemotherapy and surgery. Blood values, BMI fluctuations, and demographic information were important variables. The SHAP method was used to determine the significance of each factor. The Random Forest (RF) model performed better in external validation than other models, with an AUC of 87.4% and an F1 score of 74%.

Hwangbo et al. [16] examined the medical records of 1,002 patients with high-grade serous ovarian cancer to predict platinum sensitivity. Using stepwise selection, six key variables were identified. Of the four machine learning techniques tested, logistic regression performed best in detecting platinum-resistant patients, achieving an AUC of 0.741. Based on these findings, a web-based nomogram was developed.

Recent advancements in multi-omics integration, powered by machine learning, have transformed the landscape of personalized medicine. Multi-omics approaches combine diverse datasets, such as genomics, transcriptomics, proteomics, and metabolomics, to provide a comprehensive understanding of biological systems and disease mechanisms. Machine learning algorithms, particularly deep learning and ensemble methods, have proven effective in addressing the challenges of multi-omics data, such as high dimensionality, heterogeneity, and noise [17]. Techniques like data fusion and feature selection have enhanced the integration and interpretation of these diverse data types, allowing for the identification of clinically relevant biomarkers and predictive signatures [18]. In the context of ovarian cancer, these advancements have paved the way for individualized chemotherapy approaches by enabling the identification of molecular subtypes and patient-specific treatment responses. Recent studies have demonstrated the utility of integrating multi-omics data to predict chemotherapy sensitivity, uncover resistance mechanisms, and guide therapeutic decisions [19].

In summary, the integration of gene expression data with machine learning has shown significant promise in personalizing chemotherapy for ovarian cancer. These advancements underscore the potential for improved treatment outcomes through more precise and individualized therapeutic strategies.

Materials and Methods

This section outlines the methods applied to obtain the results. Initially, the data was preprocessed to be suitable for advanced gene selection methods and machine learning algorithms. This involved removing duplicates and formatting the data to ensure compatibility with machine learning approaches. Next, the mutual_info_classif method was used to select the most informative genes, which were then used as identifiers for training classifier techniques. Ten classifier algorithms were compared for efficacy, and the best one was selected based on its performance using four assessment metrics: accuracy (AC), precision (Pre), recall (Rec), and F1 score (Figure 1).

Datasets

A total of 58 samples were used for feature selection and machine learning techniques. The dataset was downloaded from the Gene Expression Omnibus (GEO) (ID: GSE30161) and includes data for 45,782 genes.

Preprocessing

Several crucial steps were undertaken to prepare the dataset for machine learning applications

Removing Duplicates: Duplicate records were identified and eliminated to ensure that each entry in the dataset was unique. This step helps prevent biased or skewed results during model training.

Handling Missing Data

Genes with missing data were removed, ensuring the dataset was complete and suitable for machine learning techniques.

Feature selection

The mutual_info_classif technique was employed to select the most effective features that would serve as identifiers to measure responses to Carboplatin/Taxol treatment. This method selected 20 key features. Mutual Information (MI) measures the quantity of shared information between two random variables. In gene selection, MI is used to identify a subset of genes most relevant to the target variable, such as cancer types [20]. MI has two significant advantages: it is versatile across various machine learning models and computationally efficient for feature selection. Formally, MI is defined as follows, where X represents the random variables (genes) and Y denotes the target variable (e.g., cancer types).

$$I(X,Y) = \sum\sum p(X,Y) log \frac{p(x,y)}{p(x)p(y)} \quad (1)$$

$$= H(Y) - H(Y/X) \quad (2)$$

Where H(Y—X) is the conditional entropy of Y in the case of X is known.

Machine Learning Approaches

Ten machine learning techniques were utilized to address the problem: Support Vector Machine (SVM), Decision Trees (DT), K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), Random Forest (RF), Gradient Boosting (GB), AdaBoost, Extra Trees, XGBoost, and LightGBM. Each technique has unique benefits:

Decision Trees (DT)

Known for their interpretability.

K-Nearest Neighbors (KNN)

Simple and effective for small datasets.

Multilayer Perceptron (MLP)

A deep learning method that captures complex patterns.

Random Forest (RF)

Robust due to its ensemble learning approach.

Gradient Boosting (GB)

Focuses on sequential error correction.

AdaBoost

Concentrates on difficult-to-classify cases.

Extra Trees

Increases accuracy and randomness.

XGBoost

Optimized performance with regularization.

LightGBM

Efficient and uses low memory.

This diverse set of methods was used to ensure a comprehensive assessment and determine which model performed best for the task.

Support Vector Machine

Support Vector Machines (SVMs) are primarily used for classification tasks due to their strong performance in this area, though they can also be applied to regression. SVMs work by finding an optimal hyperplane in an n-dimensional space to separate data points into distinct classes [21]. Despite their popularity, SVMs have several notable limitations:

They struggle with large datasets compared to smaller ones.

They perform poorly with noisy or overlapping target classes.

They are unsuitable for scenarios where the number of features exceeds the number of samples.

These drawbacks significantly affect their application to gene expression data, which is often noisy and has far more genes than samples [22].

K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a straightforward and effective algorithm based on the principle that similar data points are typically located close to one another. It predicts the class of a new data point by calculating its distance from previously labeled points, assigning it to the class of its nearest neighbors. This makes KNN particularly useful for applications like recommendation systems. Selecting the optimal number of neighbors (K) is crucial for achieving high accuracy, which is usually done by testing various values of K [23]. Despite its simplicity and effectiveness for small datasets, KNN can encounter challenges with noisy or missing data, inefficiency when dealing with large datasets, and struggles with high-dimensional data.

Decision Tree

A Decision Tree (DT) is a supervised machine learning method commonly used for classification tasks [23]. It works by iteratively splitting the data based on specific attributes, making it intuitive and requiring less pre-processing compared to other algorithms. However, decision trees can become overly complex, prone to overfitting, and computationally expensive, particularly when dealing with multiple class labels. The process begins with the full dataset, selecting the best attribute using an attribute selection measure (ASM), and recursively splitting the dataset into subsets based on attribute values [24].

Gradient Boosting

Gradient Boosting (GB) is a powerful ensemble

learning method that builds a strong predictive model by combining multiple weak learners, typically decision trees, in a sequential manner. In contrast to Gaussian Naïve Bayes (GNB), which is based on Bayes' theorem, GB focuses on reducing errors from previous iterations by adjusting model weights, thereby improving predictive performance. It works by assigning data points to classes based on probabilities calculated from the model's predictions. Although it is highly effective for many classification tasks, GB can be computationally expensive and sensitive to noise in the data [25, 26, 27].

Multilayer Perceptron

Multilayer Perceptron (MLP) is a robust feedforward neural network commonly applied in supervised learning tasks such as pattern recognition, classification, and prediction [21]. Its fully connected architecture allows efficient mapping of inputs to outputs. MLP employs hidden layers for complex computations and optimizes weights iteratively using backpropagation to minimize errors, making it highly effective for learning intricate patterns [28].

Random Forest

Random Forest (RF) is a versatile ensemble machine learning technique that combines multiple decision trees to improve accuracy and robustness. It is highly resilient to overfitting and effectively handles diverse data types. However, its "black-box" nature reduces interpretability, and it can be computationally intensive when dealing with large datasets [29].

AdaBoost

AdaBoost is a boosting algorithm that combines multiple weak classifiers to form a strong classifier. By iteratively adjusting weights for misclassified instances, it enhances accuracy and overall model performance. However, AdaBoost can be sensitive to noisy data, which may impact its effectiveness in certain scenarios [30].

Extra Trees

Extra Trees, or Extremely Randomized Trees, is an ensemble method that builds multiple decision trees using random subsets of data and features. It is computationally faster than traditional decision trees and helps reduce overfitting. However, it may be less accurate than other boosting methods when applied to complex datasets [31].

XGBoost

XGBoost is a powerful and highly efficient gradient boosting algorithm, known for its scalability and strong performance on structured data. It incorporates regularization to minimize overfitting and is widely used in competitive machine learning due to its speed, precision, and reliability [32].

LightGBM

LightGBM is a gradient boosting framework optimized for speed and efficiency, particularly with large datasets. It utilizes a histogram-based approach to enhance training time and memory usage, delivering high performance on large-scale data while maintaining accuracy [33].

Experiential setup

With an Intel Core™ i5-1335U processor and 8 GB RAM, Python software was utilized to implement machine learning techniques. The effectiveness of the models was rigorously evaluated using ovarian cancer as a case study. To ensure reliable results and better assess the model's generalization performance, a cross-validation method was employed to partition the datasets into training and testing groups.

A Cross-Validation

Cross-Validation (CV) is a statistical method used in ML that aims to minimize or eradicate overfitting issues in different ML approaches [34]. The k CV method allows a model to be trained on several training datasets instead of just one by training the ML algorithm on each of the k-folds created by folding the dataset. This leads to the model's ability to generalize, a sign of a strong model. It also aids in providing a clearer picture of how well the algorithmic forecast performed. Figure 2 shows how the datasets are partitioned into k-folds, such as k = 5.

Evaluation Metrics

The performance of any classification approach is typically evaluated using four key metrics. These metrics are designed to comprehensively assess a classifier's effectiveness in predicting outcomes. The evaluation criteria are as follows:

Accuracy (AC)

Accuracy is an essential evaluation metric used to identify the most successful classifier for a given dataset. In machine learning, accuracy refers to the ratio of correctly predicted observations to the total number of observations. It is calculated using the following formula [35]:

$$AC = \frac{TP + TN}{TP + FP + TN + FP} \quad (1)$$

Where:

TP (True Positive): Cases correctly predicted as positive (e.g., correctly identifying cancerous cases).

TN (True Negative): Cases correctly predicted as negative (e.g., correctly identifying non-cancerous cases).

FP (False Positive): Cases incorrectly predicted as positive (e.g., non-cancerous cases mistakenly classified as cancerous).

FN (False Negative): Cases incorrectly predicted as negative (e.g., cancerous cases mistakenly classified as non-cancerous).

recision (Pre)

Precision measures the proportion of correctly predicted positive cases out of all cases predicted as positive. It evaluates the classifier's ability to avoid false positives and is calculated as follows [35]:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall (Rec)

Recall, also referred to as sensitivity, quantifies the proportion of actual positive cases correctly identified by the model. It is particularly useful for assessing how well the classifier captures all relevant cases. The recall formula is as follows [35]:

$$Rec = \frac{TP}{TP + FN} \quad (3)$$

F1-score (F1)

The F1-score is the weighted harmonic mean of precision and recall, providing a single metric that balances both. It ranges from 0 (worst) to 1 (best) and is particularly useful when dealing with imbalanced datasets. The F1-score is calculated using the following formula [32]:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

Results

Without using feature selection

Table 1 presents the performance comparison of ten classifier algorithms for predicting chemotherapy response in ovarian cancer without applying a feature selection method. The results show that the K-Nearest Neighbors (KNN) classifier achieved the highest performance among the tested techniques.

Applying mutual_info_classif

Twenty genes were selected for analysis, including CYAT1, BBS7, AP5S1, MZT2B, BCL3, MRI8085, AS1, AF186192, PCP4L1, FAAH, RP11, PTPN20B, 574H6, LOC146880, XCL1, TMBIM6, SAFB, SLC25A14, LRMP, and LOC101929450. These genes served as identifiers to train ten different classifier methods. Table 2 summarizes the comparative performance of these classifiers in predicting chemotherapy responses in ovarian cancer. The mutual information-based feature selection method (mutual_info_classif) was employed to identify the most important genes for training. The results demonstrated that the Random Forest (RF) classifier outperformed other techniques, achieving the highest accuracy and superior evaluation metrics.

The results of the ten machine learning algorithms were compared before and after applying the feature selection technique. Table 1 shows the outcomes of employing the ten machine learning approaches without feature selection. These results indicate that the highest performance was achieved by the K-Nearest Neighbors (KNN) algorithm, which accomplished 76%, 79%, 76%, and 77% for accuracy, precision, recall, and F1-score, respectively. The lowest performance, however, was observed when applying the Multi-Layer Perceptron (MLP) and LightGBM algorithms.

In contrast, Table 2 presents the outcomes of the ten machine learning algorithms with feature selection applied. The results indicate substantial improvements in the performance of most machine learning algorithms. The

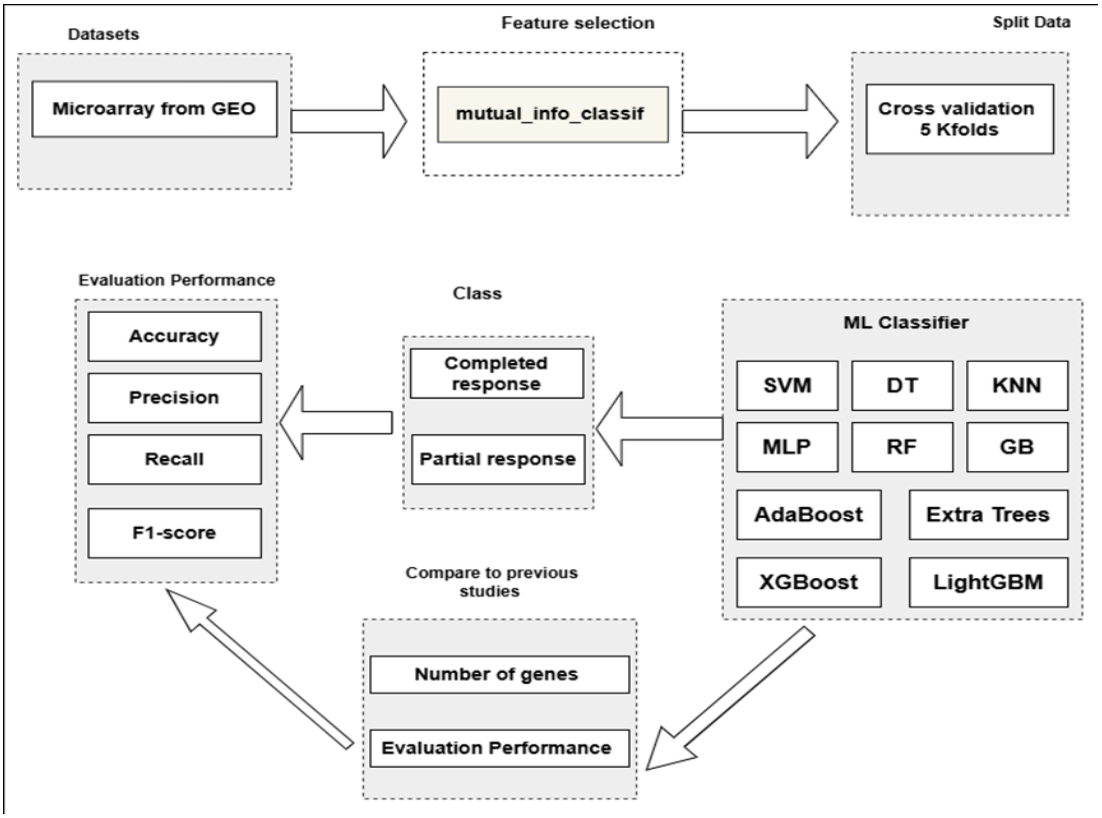

Figure 1. The System Model Structure

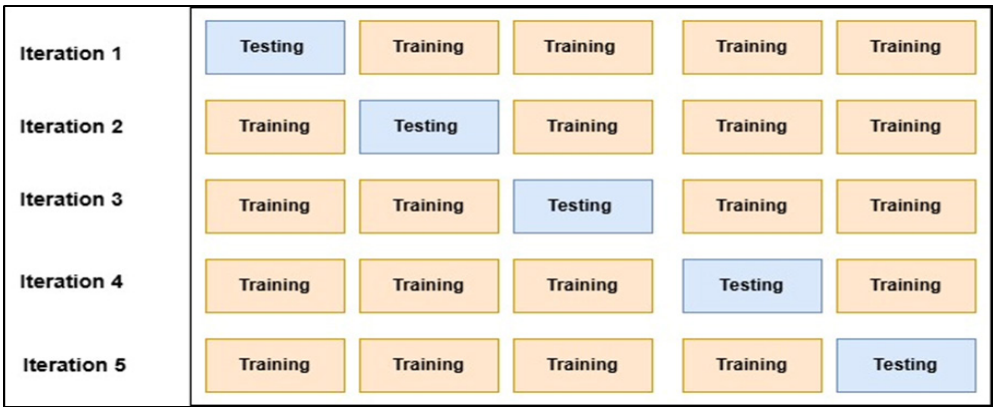

Figure 2. Cross-Validation into 5 Kfold

Table 1. Applying ML Approaches when before Applying Feature Selection Method Used

Classifier	Accuracy %	Precision %	Recall %	F1-score %
SVM	65	64	65	63
DT	65	71	65	64
KNN	76	79	76	77
MLP	59	35	59	44
RF	59	57	59	56
GB	71	75	71	71
AdaBoost	65	65	65	61
Extra Trees	76	77	76	76
XGBoost	71	71	71	71
LightGBM	59	35	59	44

highest enhancement was observed when the RF classifier was employed, achieving 97%, 98%, 97%, and 97.5% for accuracy, precision, recall, and F1-score, respectively.

The carboplatin predictor successfully distinguished between platinum-sensitive and platinum-resistant patients, achieving a statistically significant result ($p = 0.019$). Its performance parameters included a positive predictive value (PPV) of 65%, a negative predictive value (NPV) of 78%, a specificity of 33%, and a sensitivity of 93%. Similarly, the Paclitaxel predictor demonstrated effective stratification of patient responses with a sensitivity of 96%, specificity of 26%, PPV of 61%, and NPV of 86% ($p = 0.033$).

Discussion

This study employed ten machine learning approaches and gene expression patterns to determine which genes can predict a patient's response to platinum-based chemotherapy. The results demonstrated high accuracy, making the proposed model a reliable clinical indicator. Specifically, in the GEO validation sets, the Random Forest classifiers achieved 97%, 98%, 97%, and 97.5% accuracy, precision, recall, and F1-score, respectively, when feature selection (mutual_info_classif) was applied, as shown in Table 2. In contrast, the performance of most machine learning models declined significantly when

Table 2. The Outcomes of Applying ML Approaches after Using Feature Selection Method

Classifier	Accuracy %	Precision %	Recall %	F1-score %
SVM	76	76	76	76
DT	76	79	76	77
KNN	71	75	71	71
MLP	65	67	65	65
RF	97	98	97	97.5
GB	71	71	71	71
AdaBoost	76	76	76	76
Extra Trees	94	95	94	94
XGBoost	65	67	65	65
LightGBM	59	35	59	44

feature selection was omitted, as presented in Table 1. Our findings were compared with those of previous studies [36, 37, 38], which reported accuracy levels above 90% in both training and validation sets. The proposed model, achieving an accuracy of 97%, represents a substantial improvement over prior works in this field. For instance, [11] employed a Random Forest (RF) algorithm and reported an accuracy of 93%. The superior performance of the current model can be attributed to advancements in hyperparameter optimization, more effective feature selection methods, or the use of a more comprehensive and well-curated dataset. Similarly, [16] achieved an RF accuracy of 87.4%, further highlighting the performance gap. This notable improvement underscores the potential contributions of algorithmic refinements and preprocessing enhancements in the present approach.

In contrast, [20] reported an accuracy of 74.1% using stepwise logistic regression, identifying six important genes. The relatively lower performance of regression-based methods suggests limitations in capturing complex non-linear relationships in the data an issue that the proposed model effectively addresses. Furthermore, [9] achieved an accuracy of 79% using RF, further emphasizing the advancements of the current study. These improvements are likely due to the integration of a more refined RF variant, novel techniques for data preprocessing, or more sophisticated feature engineering.

Collectively, these comparisons highlight the innovative aspects of the proposed methodology, which sets a new benchmark in predictive modeling for identifying critical genes or biomarkers.

In conclusion, this study highlights the potential of using gene expression data in combination with machine learning algorithms to improve chemotherapy outcomes for ovarian cancer. By employing the mutual_info_classif method, a crucial group of 20 genes was identified as having a significant impact on treatment response. The Random Forest classifier emerged as the best-performing model, achieving high metrics in accuracy, precision, recall, and F1-score.

The results also demonstrated the effectiveness of the carboplatin predictor in distinguishing between platinum-resistant and platinum-sensitive patients. Furthermore, significant differences were observed in the survival rates between patients expected to respond to the platinum-taxane regimen and those who were not, underscoring the clinical importance of predictive modeling in chemotherapy.

In summary, this study underscores the potential of personalized chemotherapy approaches that incorporate complex genetic data to enhance treatment effectiveness for ovarian cancer patients. While the findings are promising, the proposed model requires further validation using additional datasets to improve its generalizability. Future work should focus on training the model with larger and more diverse datasets, as well as integrating advanced machine learning approaches and feature selection methods to further enhance its efficiency and robustness.

Author Contribution Statement

Mahmood S. Khalsan: Conceptualized and designed the study, implemented the machine learning techniques, and drafted the initial manuscript. Fawaz Alloosh: Critically reviewed the manuscript, provided valuable feedback, and assisted with revisions to enhance the overall quality of the paper. Ahmed S.K. Al-Khafaji: Provided critical feedback, offered constructive suggestions, and contributed to the editing process to improve the manuscript's clarity and scientific rigor. All authors have read and approved the final version of the manuscript and take full responsibility for the integrity of the work as a whole.

Acknowledgements

Funding statement

This research was supported by the Warith International Cancer Institute, which provided financial support for the study. The institute had no role in the study design, data collection, analysis, interpretation, or manuscript preparation.

Ethical issue

All analyses were conducted using publicly available datasets that were de-identified to ensure participant privacy. The use of these datasets adheres to ethical standards, and proper acknowledgments have been provided to the data sources.

Conflict of interest

There are no conflicts of interest to declare.

References

1. Zachou G, El-Khouly F, Dilley J. Evaluation of follow-up strategies for women with epithelial ovarian cancer following completion of primary treatment. Cochrane Database Syst Rev. 2023;8(8):CD006119. 10.1002/14651858.CD006119.pub4
2. Coburn SB, Bray F, Sherman ME, Trabert B. International patterns and trends in ovarian cancer incidence, overall and by histologic subtype. Int J Cancer. 2017;140(11):2451–60. 10.1002/ijc.30676.
3. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2020; 474(7353):609-615.
4. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1-22.
5. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018;15(141). https://doi.org/10.1098/rsif.2017.0387.
6. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2015;13:8-17. https://doi.org/10.1016/j.csbj.2014.11.005.
7. Khalid M, Shah AA, Hussain R. Personalized treatment of ovarian cancer using machine learning: A comprehensive review. Comput Biol Med. 2021;136:104726.
8. Jiang H, Wang X, Zhang W. Integrating multi-omics data with machine learning for predicting chemotherapy response in

ovarian cancer. Sci Rep. 2023;13(1):8561.
9. Yu KH, Levine DA, Zhang H, Chan DW, Zhang Z, Snyder M. Predicting Ovarian Cancer Patients' Clinical Response to Platinum-Based Chemotherapy by Their Tumor Proteomic Signatures. J Proteome Res. 2016;15(8):2455-65. https://doi.org/10.1021/acs.jproteome.5b01129
10. Fekete JT, Ősz Á, Pete I, Nagy GR, Vereczkey I, Győrffy B. Predictive biomarkers of platinum and taxane resistance using the transcriptomic data of 1816 ovarian cancer patients. Gynecol Oncol. 2020;156(3):654-61. https://doi.org/10.1016/j.ygyno.2020.01.006.
11. Buttarelli M, Ciucci A, Palluzzi F, Raspaglio G, Marchetti C, Perrone E, et al. Identification of a novel gene signature predicting response to first-line chemotherapy in brca wild-type high-grade serous ovarian cancer patients. J Exp Clin Cancer Res. 2022;41(1):50. https://doi.org/10.1186/s13046-022-02265-w.
12. Weberpals JI, Pugh TJ, Marco-Casanova P, Goss GD, Andrews Wright N, Rath P, et al. Tumor genomic, transcriptomic, and immune profiling characterizes differential response to first-line platinum chemotherapy in high grade serous ovarian cancer. Cancer Med. 2021;10(9):3045-58. https://doi.org/10.1002/cam4.3831.
13. Huan Q, Cheng S, Ma HF, Zhao M, Chen Y, Yuan X. Machine learning-derived identification of prognostic signature for improving prognosis and drug response in patients with ovarian cancer. J Cell Mol Med. 2024;28(1):e18021. https://doi.org/10.1111/jcmm.18021.
14. Lu TP, Kuo KT, Chen CH, Chang MC, Lin HP, Hu YH, et al. Developing a prognostic gene panel of epithelial ovarian cancer patients by a machine learning model. Cancers (Basel). 2019;11(2). https://doi.org/10.3390/cancers11020270.
15. Hsu WH, Ko AT, Weng CS, Chang CL, Jan YT, Lin JB, et al. Explainable machine learning model for predicting skeletal muscle loss during surgery and adjuvant chemotherapy in ovarian cancer. J Cachexia Sarcopenia Muscle. 2023;14(5):2044-53. https://doi.org/10.1002/jcsm.13282.
16. Hwangbo S, Kim SI, Kim JH, Eoh KJ, Lee C, Kim YT, et al. Development of machine learning models to predict platinum sensitivity of high-grade serous ovarian carcinoma. Cancers (Basel). 2021;13(8). https://doi.org/10.3390/cancers13081875.
17. Ye M, Lin Y, Pan S, Wang ZW, Zhu X. Applications of multi-omics approaches for exploring the molecular mechanism of ovarian carcinogenesis. Front Oncol. 2021;11:745808. https://doi.org/10.3389/fonc.2021.745808.
18. Khalsan M, Machado L, Al-Shamery E, Ajit S, Anthony K, Mu M, et al. A survey of machine learning approaches applied to gene expression analysis for cancer prediction. IEEE Access. 2022;10:27522-34. https://doi.org/10.1109/ACCESS.2022.3146312.
19. Chakraborty S, Sharma G, Karmakar S, Banerjee S. Multi-omics approaches in cancer biology: New era in cancer therapy. Biochim Biophys Acta Mol Basis Dis. 2024;1870(5):167120. https://doi.org/10.1016/j.bbadis.2024.167120.
20. Vergara JR, Estévez PA. A review of feature selection methods based on mutual information. Neural Comput Appl. 2014 Jan;24:175-86. https://doi.org/10.1007/s00521-013-1368-0
21. Maharjan A. Machine Learning Approach for Predicting Cancer Using Gene Expression. UNLV Theses, Dissertations, Professional Papers, and Capstones; 2020. 3922.
22. Farooq MA, Corcoran P, Rotariu C, Shariff W. Object detection in thermal spectrum for advanced driver-assistance systems (ADAS). IEEE Access. 2021;9:156465-81.
23. Cosma G, Brown D, Archer M, Pockley A. A survey on computational intelligence approaches for predictive modeling in prostate cancer. Expert Syst Appl. 2016;70. https://doi.org/10.1016/j.eswa.2016.11.006.
24. Zhong Y. The analysis of cases based on decision tree. In2016 7th IEEE international conference on software engineering and service science (ICSESS) 2016 Aug 26 (pp. 142-147). IEEE.
25. Hartatik Uns H, Purnomo A, Hartono R, Munawaroh H. Naïve bayes approach for expert system design of children skin identification based on android. IOP Conference Series: Materials Science and Engineering. 2018;333:012105. https://doi.org/10.1088/1757-899X/333/1/012105.
26. Wu X, Kumar V, Quinlan R, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. Knowl Inf Syst. 2007;14. https://doi.org/10.1007/s10115-007-0114-2
27. Raizada RD, Lee YS. Smoothness without smoothing: Why gaussian naive bayes is not naive for multi-subject searchlight studies. PLoS One. 2013;8(7):e69566. https://doi.org/10.1371/journal.pone.0069566.
28. Xie R, Wen J, Quitadamo A, Cheng J, Shi X. A deep auto-encoder model for gene expression prediction. BMC Genomics. 2017;18(Suppl 9):845. https://doi.org/10.1186/s12864-017-4226-0
29. Quist J, Taylor L, Staaf J, Grigoriadis A. Random forest modelling of high-dimensional mixed-type data for breast cancer classification. Cancers (Basel). 2021;13(5). https://doi.org/10.3390/cancers13050991.
30. Zhang Y, Ni M, Zhang C, Liang S, Fang S, Li R, Tan Z. Research and application of AdaBoost algorithm based on SVM. In2019 IEEE 8th joint international information technology and artificial intelligence conference (ITAIC) 2019 May 24 (pp. 662-666). IEEE
31. Ahmad M, Reynolds J, Rezgui Y. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. J Clean Prod. 2018;203:810-21. https://doi.org/10.1016/j.jclepro.2018.08.207.
32. Chen M, Liu Q, Chen S, Liu Y, Zhang CH, Liu R. Xgboost-based algorithm interpretation and application on post-fault transient stability status prediction of power system. IEEE Access. 2019;7:13149-58. https://doi.org/10.1109/ACCESS.2019.2893448.
33. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems. 2017;30.
34. Khalsan M, Mu M, Al-shamery ES, Ajit S, Machado L, Agyeman MO. Developing a Multidimensional Fuzzy Deep Learning for Cancer Classification using Gene Expression Data. In9th International Conference on Computer Science, Engineering and Applications 2023 Dec 20 (pp. 37-49). Academy and Industry Research Collaboration Center (AIRCC).
35. Khalsan M, Mu M, Al-Shamery ES, Ajit S, Machado LR, Agyeman MO. A novel fuzzy classifier model for cancer classification using gene expression data. IEEE Access. 2023 Oct 17;11:115161-78.
36. Amniouel S, Yalamanchili K, Sankararaman S, Jafri MS. Evaluating ovarian cancer chemotherapy response using gene expression data and machine learning. BioMedInformatics. 2024;4(2):1396-424. https://doi.org/10.3390/biomedinformatics4020077.
37. Gonzalez Bosquet J, Devor EJ, Newtson AM, Smith BJ, Bender DP, Goodheart MJ, et al. Creation and validation of models to predict response to primary treatment in serous ovarian cancer. Sci Rep. 2021;11(1):5957. https://doi.

org/10.1038/s41598-021-85256-9.
38. Ferriss JS, Kim Y, Duska L, Birrer M, Levine DA, Moskaluk C, et al. Multi-gene expression predictors of single drug responses to adjuvant chemotherapy in ovarian carcinoma: Predicting platinum resistance. PLoS One. 2012;7(2):e30550. https://doi.org/10.1371/journal.pone.0030550.