MINISTRY OF SCIENCE AND HIGHER EDUCATION OF THE RUSSIAN FEDERATION
Federal State Autonomous Educational Institution of Higher Education
Peter the Great St. Petersburg Polytechnic University

Institute of Computer Science and Cybersecurity

Higher School of Artificial Intelligence Technology

Direction 02.03.01 Mathematics and computer Science

Literature Review

Machine learning approaches for assessing drug resistance in cancer treatment

Student,
group 5130201/20102 _____ Tishenko A. A.

Supervisor, Ph. D. _____ Motorin D. E.

«_____»_____ 2024г.

Saint-Petersburg, 2024

Keywords

Cancer treatment, drug resistance, chemotherapy resistance, machine learning, artificial intelligence, interpretable machine learning.

Introduction

Cancer remains one of the leading causes of mortality worldwide, presenting a significant challenge to global health despite considerable advancements in medical research and treatment. According to the data from the International Agency for Research on Cancer (IARC), there were over 19 million new cancer cases and nearly 10 million cancer-related deaths globally in 2020 [11]. The high mortality rate is attributed to factors such as late diagnosis, the aggressive nature of certain cancer types, and the complexity of developing effective treatment strategies. The impact of cancer is profound, not only on the patients but also on their families and healthcare systems, underscoring the need for continual improvement in cancer management.

Chemotherapy is a cornerstone in cancer treatment, widely utilized either as a primary or adjuvant therapy to target and destroy rapidly dividing cancer cells. It is often the first-line treatment for various cancers, including ovarian, lung, and cervical cancers [12]. Standard chemotherapy regimens, such as platinum-based drugs combined with paclitaxel, have significantly improved patient survival rates by inhibiting tumor growth and controlling metastasis. For example, in advanced ovarian cancer, debulking surgery combined with six cycles of platinum and paclitaxel chemotherapy is the standard approach, aiming to reduce tumor burden and eradicate micrometastases [13].

However, a major challenge in chemotherapy is the development of drug resistance, which leads to treatment failure, disease progression, and adversely affects patient survival [14]. Drug resistance can be intrinsic, where tumors inherently do not respond to chemotherapy, or acquired, developing after initial responsiveness due to factors like genetic mutations and cellular adaptations. For example, in epithelial ovarian cancer (EOC), approximately 80% of patients experience relapse after initial remission because of chemotherapy resistance, making it a significant factor contributing to the ineffectiveness of treatment and the leading cause of death in these cases [1]. Similarly, in cervical cancer, up to 40% of patients exhibit resistance to platinum-based neoadjuvant chemotherapy, potentially delaying effective surgical interventions and accelerating disease progression [8]. Existing chemotherapy resistance assessment methods often have limitations such as low success rates in modeling, high costs, and time-consuming processes.

Machine learning and deep learning has made dramatic breakthroughs in recent years to contribute to the field of medicine [15]. It has been widely applied to various classification and regression problems, especially in the field of biology where the amount and complexity of data is growing. Machine learning in drug resistance is used for identifying critical genetic, molecular, and morphological features associated with resistance, developing predictive models to classify sensitive and resistant cells, and analyzing complex datasets to uncover patterns. It is also applied to design personalized treatment strategies, predict patient-specific drug responses, and explore mechanisms driving resistance, aiding in the optimization of therapeutic approaches.

Machine learning models are often considered "black boxes", as there is typically a limited understanding of how the outputs are produced from the model input. However, in clinical medicine tasks, it is equally important to understand the features that drive a model's decision-making. To address this, researchers use different tools and approaches to perform feature importance analysis, e.g., SHapley Additive exPlanations (SHAP) [16], least absolute shrinkage and selection operator (LASSO) [17], DALEX [24], and other. Such analysis helps to better understand the underlying causes and mechanisms of re-

sistance, providing valuable insights that can guide the development of more effective treatment strategies and targeted therapeutic interventions.

This article provides a comprehensive overview of how machine learning algorithms have been applied to predict and analyze chemotherapy resistance across various cancer types. We explore diverse approaches employed in the field, ranging from feature extraction and classification to regression and prognostic modeling. The goal of this review is to shed light on the current state of the art in using machine learning for assessing drug resistance and to identify opportunities for future research that can enhance the precision and effectiveness of cancer treatment strategies.

1 Machine learning and chemotherapy resistance

Machine learning has been widely applied to various classification, regression, feature extraction and many other problems in the field of biology and medicine. The field of cancer treatment has also not been left aside, in particular, machine learning has recently been actively used in research related to the problem of cancer cell chemotherapy resistance.

Authors of [1] applied and compared five different machine learning algorithms to classify cancer cells based on their level of drug resistance. They extracted 112 morphological features from dataset of nearly 3000 single-cell quantitative phase images of epithelial ovarian cancer (EOC) cells. After that, authors employed five supervised machine learning algorithms, Tree, Naive Bayes, K-nearest neighbors (KNN), support vector machine (SVM), and neural network (NN), to perform multi-classification on four types of drug-resistant cancer cells. The optimal classification algorithm was determined by comparing the classification testing accuracy for each cell type and the confusion matrix. The chosen trained model was then used for further interpretable analysis.

Another study aims to evaluate the potential of mitochondria-related chemoradiotherapy (CRT) resistance (MRCRTR) genes in predicting esophageal cancer prognosis using machine learning [3]. Authors used machine learning algorithms for both classification and regression tasks. For classification they applied seven algorithms: generalized linear model (GLM), K-nearest neighbor (KNN), least absolute shrinkage and selection operator (LASSO) regression, neural network (NN), random forest (RF), support vector machine (SVM), extreme gradient boosting (XGB). They applied those algorithms to pretty similiar task as in [1], but in this paper authors identified only two classes – CRT response and CRT non-response. The authors did not stop at classification alone, but also trained 10 machine learning algorithms, including random survival forest (RSF), elastic network (Enet), LASSO, ridge, stepwise Cox, Coxboost, partial least squares regression for Cox (plsRcox), supervised principal components (SuperPC), generalized boosted regression modeling (GBM), and survival support vector machine (survival-SVM), to build consensus prognostic model to predict MRCRTR score. Using the leave-one-out cross-validation (LOOCV) framework, a total of 101 algorithm combinations were applied to match prognostic models.

Machine learning algorithms also was successfully applied for same classification task as in [1] and [3] by authors of [4]. They employed robust machine learning algorithm based on principal component analysis and linear discriminant analysis (PCA-LDA) to extract the feature of blood-SERS data and establish an effective predictive model for identifying the radiotherapy resistance subjects from sensitivity ones, and for identifying the nasopharyngeal cancer (NPC) subjects from healthy ones.

The authors of article [2] chose a different approach by applying machine learning algorithms from the specialized software CellProfiler [18] to extract quantitative image features. They subsequently used bioinformatics analysis to explore the relationship between these features of intra-tumor heterogeneity (ITH) and drug resistance. Notably, the authors did not aim to train new models but instead utilized pre-trained algorithms from CellProfiler. Unlike studies [1], [3], and [4], where algorithms were employed for regression and classification tasks, this research focused specifically on extracting quan-

titative features from images. Based on CellProfiler, the authors constructed a pipeline for the extraction and analysis of these features, which enabled them to draw conclusions regarding the connection between these features and drug resistance in cancer cells.

In [5], the authors performed differential protein analysis on the expression profiles of 745 proteins related to platinum-based chemotherapy resistance. They used LASSO regression [17] to select 10 proteins linked to chemotherapy outcomes, followed by univariate logistic regression on nine clinical factors. Variables with $p < 0.1$ were included in a multivariate logistic regression analysis, resulting in four significant variables: three proteins and one clinical parameter (postoperative residual tumor). This analysis enabled the construction of a predictive machine-learning model for chemotherapy resistance in patients with EOC.

The authors of article [6] applied machine learning algorithms for two goals. Firstly, they used algorithms to extract genes highly related with therapy resistance. Each sample of their data contained the expression of 8687 genes and only a small portion was correlated with targeted therapy resistance. To extract highly related genes in this study authors attempted seven algorithms, including Least Absolute Shrinkage and Selection Operator (LASSO), Light Gradient Boosting Machine (LightGBM), Monte Carlo Feature Selection (MCFS), Minimum Redundancy Maximum Relevance (mRMR), Random Forest (RF) -based, Categorical Boosting (CATBoost), and eXtreme Gradient Boosting (XGBoost). Secondly, they selected four algorithms to perform binary classification (resistant vs sensitive) of tumor cells based on extracted features, namely, random forest (RF), support vector machine (SVM), K-Nearest Neighbors (KNN), and decision tree (DT).

The authors of article [7] took an alternative approach: instead of directly predicting chemotherapy resistance, they constructed the machine learning-derived immunosenescence-related score (MLIRS) score. Patients with high MLIRS scores had a worse prognosis. In contrast, the low MLIRS score group demonstrated greater sensitivity to both chemotherapy and immunotherapy. To obtain an optimal hazard scoring system, they trained a total of 101 combined machine learning algorithms (based on 10-fold cross-validation) across 10 basal categories: survival support vector machine (survival-SVM), CoxBoost, random survival forest (RSF), Lasso, stepwise Cox, partial least squares regression for Cox (plsRcox), Ridge, supervised principal components (SuperPC), elastic network (Enet), and generalized boosted regression modeling (GBM). In this study, these algorithms were applied to a regression task, allowing the authors to compute the coefficients for the MLIRS formula:

$$\text{MLIRS} = (\text{expr}_{\text{gene1}} \times \text{coff}_{\text{gene1}}) + (\text{expr}_{\text{gene2}} \times \text{coff}_{\text{gene2}}) + \ldots + (\text{expr}_{\text{gene}_n} \times \text{coff}_{\text{gene}_n})$$

where: $\text{expr}_{\text{gene}}$ denotes the expression level of each gene, $\text{coff}_{\text{gene}}$ represents the coefficient for each gene, as determined by the model. Authors derived the C-index value of each machine learning algorithm in each dataset and identified the algorithm with the largest mean C-index as the optimal hazard scoring algorithm.

The authors of [9] developed a machine learning model to predict cisplatin sensitivity based on gene expression changes induced by cisplatin treatment. They combined gene expression data from sensitive ovarian cancer cell lines and patients with specific signaling alterations to identify a gene signature. Using this signature, they trained TabNet, an interpretable deep learning algorithm for tabular data, to perform binary classification of sensitivity to cisplatin. Also several other machine learning algorithms, including Ridge, LASSO, Elastic Net, Nu-Support Vector Classification (Nu-SVC), XGBoost, and Random

Forest, were applied to the same task for comparission with TabNet.

Same as in the [2], the authors of [10] used algorithms from the specialized software called Acapella (developed by PerkinElmer [25]) to extract 624 quantitative image features from cellular images. These quantified features were then analyzed using deep learning in order to reduce the dimensionality of the data. The authors ran 46 separate deep learning models, all with the same input data but with different numbers of output nodes in order to reduce the dimensionality of the data to greater and lesser extents. Then they scored these 46 different models based on their ability to recreate the input data. After that they selected the model with the lowest number of dimensions, in this case 27, that reached reconstruction error plateau. This deep learning model identified a continuous 27-dimension space describing all of the observed cell morphologies. Upon this model a random forest classifier was trained on populations of cells labeled as either drug sensitive or drug resistant.

2 Datasets

Data plays a crucial role in machine learning, serving as the foundation for model training and evaluation. The quality and quantity of data directly influence the performance and generalizability of machine learning algorithms. In the fields of biology and medicine, data collection is often costly and time-consuming. Additionally, the complexity and variability inherent in biological systems further complicate data acquisition and interpretation. In cancer research, these challenges are even more pronounced due to the heterogeneity of tumors and the intricate nature of cancer biology. However, there are valuable resources available, such as the Gene Expression Omnibus (GEO) database [20] and The Cancer Genome Atlas (TCGA) database [19], which provide researchers with access to extensive datasets. Moreover, nonprofit organizations like the American Type Culture Collection (ATCC) [22] enable researchers to obtain biological materials, including cancer cells.

In articles [1], [4], [8] and [10] authors decided to prepare their own datasets specifically for their research.

In [1] four kinds of epithelial ovarian cancer cells with different drug sensitivity (SKOV3, SKOV3_Ta_2μM, SKOV3_Ta_8μM, and SKOV3_Ta_20μM) were studied. The SKOV3 cells were sourced from the ATCC [22] and preserved at the Obstetrics and Gynecology Laboratory of Peking University People's Hospital. The drug-resistant characteristics of SKOV3_Ta_2μM, SKOV3_Ta_8μM, and SKOV3_Ta_20μM were acquired by progressively exposing SKOV3 cells to varying concentrations of paclitaxel. After approximately ten months, all the drug-resistant cancer cells were acquired. They then utilized Digital Holographic Flow Cytometry (DHFC), an advanced technology for label-free, high-throughput cell detection. Using DHFC along with additional post-processing, the authors generated a dataset comprising approximately 3000 a quantitative phase images (QPIs) of EOC cells, each sized at 300 by 300 pixels. Fig. 1 presents the reconstructed QPIs of EOC cells with various degrees of drug resistance.

The dataset in [4] was based on clinical plasma samples from 60 healthy volunteers which were used as a control group, and 60 nasopharyngeal cancer patients (30 plasma samples from radiotherapy sensitivity patients and 30 plasma samples from radiotherapy

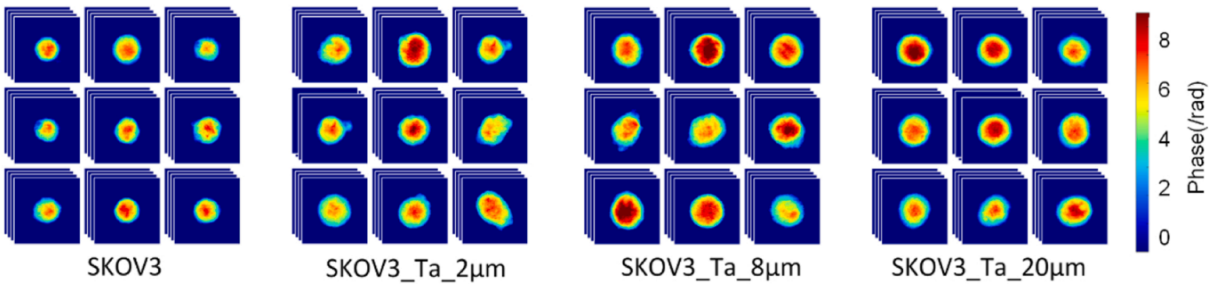

Figure 1. Reconstructed QPIs of EOC cells used by authors of [1].

resistance patients). All plasma samples were obtained from Fujian Provincial Cancer Hospital. As well as in [1], authors used unique method called surface enhanced Raman spectroscopy (SERS) to extract molecular profiles of patients plasma. Authors even claim that SERS based on surface plasmon resonance was used for this task for the first time. The SERS spectra were processed by deducting the fluorescence background signal using a fifth-order polynomial fitting method, and then the SERS signals were peak normalized, after which the spectra of the same plasma sample were averaged to represent the final SERS data for that sample.

In [8], authors prepared dataset with 259 samples. They choosed 259 patients at the People's Hospital of Gansu Province and the First and Second Hospital of Lanzhou University who were diagnosted with locally advanced cervical cancer (LACC), applied neoadjuvant chemotherapy (NACT) to them and extracted their whole blood genomic DNA. After that 24 SNPs from PTEN/PI3K/AKT pathway: PTEN, PIK3CA, Akt1, and Akt2 were selected. 70 features were generated from 24 SNPs in the raw data using the one-hot encoding method resulting in 259x70 dataset. Clinical examination, colposcopy, and abdominal computer tomography were used to estimate the change of tumor size in all patients before and after each NACT cycle. In this study, patients with a complete response and partial response were classified as NACT effective group, and patients with stable disease and progressive disease were considered NACT ineffective group.

The dataset in [10] was based on 12 drug-resistant clones, i.e., populations of cells derived from a single progenitor cell and genetically identical to it, generated from five human cancer cell lines (tongue, lung, breast, and esophageal cancers). The authors used similiar approach as in [1] to get cells with drug resistance. Clones were made resistant to cetuximab, pertuzumab, or trastuzumab, which are inhibitory antibodies targeting members of the ErbB family of receptor tyrosine kinases, over a period of six months. Following this, cells were treated with small interfering RNAs (siRNAs) targeting 536 protein kinases and then one of 11 different ErbB-inhibiting antibody drugs. In total, the study imaged 848,802,073 cells. From each cell, 624 features were extracted using Acapella image analysis software, resulting in a dataset containing 529,652,493,552 data points. This approach enabled the comparison of the effects of inhibiting ErbB kinase signaling on cell morphology in drug-sensitive and drug-resistant cancer cell lines.

Authors of articles [2], [3], [6], [7] and [9] turned to open databases to prepare datasets for their research. Authors of [2] downloaded frozen histopathologic images of 494 ovarian and 70 paracarcinoma tissues with hematoxylin–eosin (HE) staining from TCGA [19]. The corresponding clinical information, genomics, and transcriptomics profiles required for this study were also obtained from this database. Authors of [3] also used TCGA.

They downloaded information on 183 esophageal cancer patients (95 squamous cell carcinomas and 88 adenocarcinomas) was obtained, including mRNA expression profiles, clinical features such as survival time and status, age, gender, and pathological stage (T, N, and M). Additionally authors used Gene Expression Omnibus (GEO) database [20]. RNA sequencing (RNA-seq) for GSE45670 was downloaded from it. GSE45670 includes a total of 17 esophageal squamous cell carcinomas (ESCC) that did not respond to preoperative CRT, 11 ESCC that responded to preoperative CRT, and 10 samples from normal esophageal epithelium. The GEO dataset GSE53625 comprises 358 samples, including 179 ESCC tissue samples and an equal number of samples of adjacent normal tissues, along with detailed clinical data for the 179 ESCC patients. The GEO dataset GSE19417 contains data from 76 esophageal adenocarcinoma patients, offering detailed clinical data for 48 of these patients. Authors of [6] also took gene expression profile data from GEO database, specifically from accession number GSE137912. Their analysis involved 7612 samples treated with KRAS G12C inhibitors. Among these samples, 4297 were tumor cells that persisted in proliferation, whereas 3315 were tumor cells that had ceased proliferating. Each sample contained the expression of 8687 genes. In [7], authors used datasets from both TCGA and GEO and also from European Genome-Phenome Archive (EGA) [21]. Authors of [9] used GSE47856, GSE15622 and GSE146965 from the GEO database and RNAseq data from TCGA.

In article [5], authors prepared their own dataset and also used open databases. In this study, 4D data-independent acquisition (DIA) proteomic sequencing was performed on tissue-derived extracellular vesicles (tsEVs) obtained from 58 platinum-sensitive and 30 platinum-resistant patients with EOC. Also authors used the GSE15372, GSE33482, GSE26712 and GSE63885 microarray datasets from the Gene Expression Omnibus database [20]. GSE15372 and GSE33482 represent EOC cell line-derived RNA microarray datasets, comprising 5 and 5 and 6 and 6 platinum-sensitive and resistant cell line samples, respectively. GSE26712 and GSE63885 involve clinical and sequencing data for 195 and 101 EOC patients, respectively. Additionally, transcriptomic sequencing data and clinical information from the tumour tissues of 379 patients with EOC, sourced from the TCGA database [19], was used.

3 Feature importance analysis

Machine learning algorithms are often regarded as black boxes, providing powerful predictions but limited insight into the underlying mechanisms driving those predictions. In fields such as medicine and biology, however, interpretability is not just a desirable feature—it is a necessity. Transparent models and interpretable outputs are critical for ensuring that predictions and recommendations can be explained, validated, and trusted, especially when they influence life-altering decisions. In the context of cancer treatment, this need for interpretability becomes even more pressing. Misguided treatment decisions can delay the administration of effective therapies, significantly increasing the risk of mortality. Beyond guiding clinical decisions, the interpretability of features in machine learning models also offers a unique opportunity to deepen our understanding of drug resistance in cancer. By unraveling the biological and molecular underpinnings of resistance, we can develop more targeted and effective therapeutic strategies.

The authors of [1] applied the SHapley Additive exPlanations (SHAP) [16], a model

interpretation framework, to quantify and rank the feature contributions for classification models. After analysing feature importances authors were able to reduce count of the features from 112 to only 25 and even increase the accuracy of their models. Interestingly that their models were able to achieve approximately 78% even when only the top three features were utilized for classification.

In [9] authors also used SHAP to performe feature analysis. In this study, feature importance analysis was utilized to identify key genes associated with cisplatin resistance. By leveraging feature importance derived from multiple predictive models, the authors highlighted BCL2L1 as a critical gene mediating cisplatin resistance. Notably, earlier studies suggested that β-catenin expression is necessary for maintaining elevated levels of BCL2L1 in cisplatin-resistant cells. Additionally, the analysis revealed the involvement of PLK2, whose expression was linked to β-catenin activity. The authors demonstrated that lower expression levels of BCL2L1 and PLK2 correlated with improved outcomes in ovarian cancer, and inhibitors targeting these genes showed a synergistic effect when combined with cisplatin. These findings underscore the importance of the β-catenin/BCL2L1 axis in driving resistance and provide potential targets for enhancing cisplatin efficacy.

In study [2], the authors employed feature importance analysis to identify key histopathological features associated with intratumoral heterogeneity (ITH), drug resistance, and prognosis in ovarian cancer. The analysis was conducted using the R programming environment [23], leveraging tools such as the "limma" package for differential analysis and the "glmnet" package for LASSO regression modeling. 924 features were identified as differentially expressed between cancerous and non-cancerous tissues. Of these, 394 features were associated with overall survival, and 26 key features were identified at the intersection of survival analysis and differential expression.

The authors of [3] also used R programming environment to performe feature importance analysis to identify key predictor genes for mitochondrial-related CRT resistance (MRCRTR). They used the DALEX [24], an R package for model interpretability, to analyze feature importance and residual distribution, which helps interpret how different features influence model predictions. This tool provided insights into the contribution of each predictor gene across the machine learning models. The top 12 genes identified through this analysis were selected as MRCRTR predictor genes, contributing to the development of a prognostic model for esophageal cancer.

In [6] feature importance analysis was employed to identify genes associated with resistance to KRAS G12C inhibitor treatment in cancer cells. The authors used seven different feature ranking algorithms: LASSO, LightGBM, MCFS, mRMR, RF-based, CATBoost, and XGBoost. These algorithms generated feature lists based on different principles, enabling a comprehensive evaluation of gene significance. To refine the feature selection, the authors applied Incremental Feature Selection (IFS), testing the performance of classifiers like Decision Tree (DT), k-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machine (SVM) on the ranked features. By doing feature analysis they were able to highlight several key genes, such as H2AFZ, CKS1B, and TUBA1B, which were consistently ranked highly across multiple algorithms and are linked to tumor progression and drug resistance.

In study [7], univariate Cox regression analysis was employed to identify immunosenescence-related genes with prognostic significance in pancreatic cancer. The Cox proportional hazards model was used to assess the relationship between the expression of

individual genes and overall survival. The hazard ratio (HR) for each gene was estimated, with a p-value of less than 0.01 indicating statistical significance. Genes with a p-value below this threshold were selected as meaningful features for subsequent analysis, as they were considered to have a potential impact on the prognosis of pancreatic cancer patients. This approach allows for the identification of genes that might serve as independent prognostic biomarkers.

The authors of [8] used feature importance analysis based on the Random Forest (RF) model to identify key SNPs related to NACT sensitivity in LACC patients. The importance of each feature was calculated by assessing its impact on impurity reduction at each node in the RF model, with a larger decrease in impurity indicating greater feature importance. The mean decrease in impurity (MDI) was calculated using the total decrease in impurity averaged over all decision trees. The impurity g of a split was computed as:

$$g = 2 \cdot p_A \cdot p_B$$

where p_A and p_B represent the probabilities of class A and class B, respectively. The overall impurity G for a split was calculated as the weighted average of the impurities of its two sub-splits:

$$G = P_1 \cdot g_1 + P_2 \cdot g_2$$

where P_1 and P_2 are the proportions of data in the sub-splits and g_1, g_2 are the impurities of each sub-split. Feature importance for each SNP was calculated by summing the importance of all features generated by each SNP (after one-hot encoding):

$$L_i = \sum_j f_{i,j}$$

where L_i is the importance of SNP i, and $f_{i,j}$ represents the importance of feature j generated by SNP i.

4 Results

In all works, the construction of machine learning models is essentially a secondary result. First of all, studies show the applicability of these methods to tasks related to the problems of cancer cell resistance to chemotherapy. Also, using machine learning methods, the authors test their hypotheses, confirm or discover links between various characteristics of cancer cells, patient clinical data and drug resistance.

In articles [1], [4], [5], [6], [8], [9], [10], the authors try to solve the problem of determining drug resistance directly. In [4], [5], [6], [8], [9], [10], the problem of binary classification (drug resistant vs drug sensitive) is solved, and in [1], cells are classified into 4 classes, which constitute a gradation of the level of resistance of cancer cells to chemotherapy.

In [1], five different machine learning algorithms were compared, the best results were achieved using support vector machine (accuracy of 93.4%) and neural network (accuracy

of 94.5%). The classification was based on morphological features and, by constructing effective classifiers, the authors demonstrated that these features are directly related to the level of resistance of cancer cells to chemotherapy. Also, using SHapley Additive exPlanations authors showed that only a 25 of 112 features are really important for the classification.

The authors of [4], applied robust machine learning algorithm based on principal component analysis and linear discriminant analysis and established an effective predictive model with the accuracy of 96.7% for identifying the radiotherapy resistance subjects from sensitivity ones, and 100% for identifying the NPC subjects from healthy ones. Also authors showed the importance of the separation of plasma into upper and lower plasma by comparing model results, e. g. for upper plasma and radiotherapy resistance vs. radiotherapy sensitivity classification task their model achieved 98.7% accuracy while for lower plasma it is only at level of 93.9%.

LASSO-based classifier was built by authors of [5]. Their model achieved Area Under Curve (AUC) of 0.864. By analysing their model and its results authors found that three immune-related proteins—CCR1, IGHV3-35, and CD72—along with the presence of postoperative residual tumors, are strong predictors of platinum resistance in EOC patients.

In [6], authors firstly applied machine learning algorithms to extract most important features and created seven feature lists, after that they applied four classification algorithms. Their best result was achieved with CATBoost feature list and support vector machine as classification algorithms (accuracy of 93.1%). Also after analysing recieved feature lists authors were able to identify top genes associated with tumor progression and drug resistance (H2AFZ, CKS1B, TUBA1B, RRM2, BIRC5).

The study [8] employed a Random Forest model utilizing genomic features. The model successfully predicted the response to platinum-based neoadjuvant chemotherapy in patients with locally advanced cervical cancer (LACC). However, the main focus of the study was not on building the model but on analyzing feature importance to identify key genes associated with chemoresistance. Through importance analysis, the authors identified that the top three significant single nucleotide polymorphisms (SNPs)—rs4558508, rs1130233, and rs7259541—were all located within the Akt gene family. Specifically, patients carrying the heterozygous GA genotype in Akt2 rs4558508 had a significantly increased risk of chemoresistance compared to those with GG or AA genotypes.

The authors of [9] developed a deep learning model using the TabNet algorithm to predict cisplatin sensitivity based on cisplatin-perturbed gene expression data. Their model achieved over 80% accuracy, surpassing a variety of other machine learning algorithms such as ridge regression, lasso, elastic net, Nu-SVC, XGBoost, and random forest. The TabNet model consistently demonstrated strong predictive performance with an average AUC of 0.808 across 500 different sample splits. By analyzing feature importance, the authors identified several key genes contributing to cisplatin resistance, most notably BCL2L1. The upregulation of BCL2L1, along with genes like CCND1 and PLK2, was associated with poor survival in ovarian cancer patients, highlighting potential targets for overcoming drug resistance. These findings are in line with the results of [6], where important genes associated with tumor progression and drug resistance were also identified using machine learning feature selection techniques.

In [10] authors demonstated how deep learning of cell morphologies can be used to

successfully predict drug resistance state in cancer cell lines from diverse tissues. They built a classifier based on deep neural network and random forest which can identify cancer cells resistance to ErbB-family drugs with an accuracy of 74%.

In articles [2], [3], [7], the authors used machine learning for a different tasks. In [2] used machine learning algorithms from the specialized software CellProfiler [18] to extract quantitative image features and then performed statistical analysis of feature importance. The authors of [3] and [7] applied machine learning algorithms for the regression task and proposed their own scores, mitochondria related chemoradiotherapy resistance (MRCRTR) score and machine learning-derived immunosenescence-related score (MLIRS), respectively.

The study [2] demonstrated that specific computational pathomic signatures extracted from histopathological images can effectively predict drug resistance in ovarian cancer patients. By analyzing 1212 statistical image features derived from whole-slide images, the authors identified 26 key features related to patient survival. Among these, the Perimeter.sd feature, which measures the standard deviation of nuclear perimeter, stood out as the most significant predictor. A higher Perimeter.sd value was positively correlated with increased intra-tumor heterogeneity and was associated with a higher risk of platinum-based chemotherapy resistance.

The authors of [3] developed a prognostic model based on mitochondria-related chemoradiotherapy resistance (MRCRTR) genes to predict survival outcomes in esophageal cancer patients. They identified six key genes (CTSL, TBL1X, CLN8, MMP1, PDPN, and MRPL37) that have high diagnostic value for chemoradiotherapy resistance. The MRCRTR score derived from these genes showed that patients with high scores had significantly lower survival rates than those with low scores (log-rank test, $p < 0.001$). Cox regression analyses confirmed the MRCRTR score as an independent prognostic factor. Additionally, the MRCRTR score was significantly correlated with increased expression of immune checkpoints and higher angiogenesis, epithelial-mesenchymal transition (EMT), and cancer-associated fibroblast (CAF) scores.

The authors of [7] identified two immunosenescence-associated phenotypes (IMSP1 and IMSP2) with significant differences in prognosis and immune cell infiltration. The authors constructed a Machine-Learning Immunosenescence-Related Scoring (MLIRS) system using a combination of stepwise Cox regression and generalized boosted regression modeling (GBM), integrating multiple machine learning algorithms across 101 cross-validation methods. Their MLIRS model demonstrated robust prognostic performance with an Area Under Curve (AUC) of 0.91. They found that patients with high MLIRS scores had worse prognosis and lower abundance of immune cell infiltration, whereas those with low MLIRS scores showed better sensitivity to chemotherapy and immunotherapy.

Conclusion

In conclusion, machine learning approaches have emerged as pivotal tools in the assessment of drug resistance in cancer treatment, addressing one of the most formidable challenges in oncology. The integration of machine learning algorithms into cancer research and clinical practice has facilitated the identification of critical genetic, molecular, and morphological features associated with drug resistance. Reviewed studies have shown that machine learning models can accurately classify and predict resistance patterns, enabling earlier intervention and more personalized therapeutic strategies. The authors applied machine learning to analyze diverse types of data, from gene expression profiles to histopathological images. Techniques such as SHapley Additive exPlanations (SHAP) [16] and least absolute shrinkage and selection operator (LASSO) [17] have improved the interpretability of the models, transforming them from "black boxes" into transparent systems that provide meaningful insights into the mechanisms underlying drug resistance. This transparency is crucial for clinical applications, as it allows to understand and trust the model's predictions, ultimately guiding more informed decision-making processes.

Despite the significant progress, several challenges remain. The complexity of cancer biology necessitates the continual refinement of the algorithms to improve their accuracy and generalizability across different cancer types and patient populations, e. g. most of the reviewed studies focus on Chinese populations. Additionally, the integration of machine learning tools into clinical workflows requires robust validation and the establishment of standardized protocols to ensure consistency and reliability in diverse healthcare settings. Future research should focus on collecting more data, making models easier to understand, and encouraging collaboration between different fields to effectively apply computational advances in clinical scenarios.

Overall, the application of machine learning in assessing drug resistance represents a novel approach in cancer treatment, offering lots of opportunities to enhance the precision and effectiveness of therapies. By continuing to advance machine learning algorithms and support their integration into clinical practice, the medical community can significantly improve the management of drug-resistant cancers, ultimately reducing mortality rates and improving the quality of life for patients worldwide.

Table 1. Methods used in research papers. Abbreviations: Epithelial Ovarian Cancer (EOC), ESophageal Cancer (ESC), NaSopharyngeal Cancer (NSC), Lung Cancer (LC), Pancreatic Cancer (PC), Cervical Cancer (CC), Tongue Cancer (TG), Breast Cancer (BC), Decision Tree (DT), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Neural Network (NN), Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest (RF), Principal Component Analysis - Linear Discriminant Analysis (PCA-LDA), eXtreme Gradient Boosting (XGB), Generalized Linear Model (GLM), Logistic Regression (LR), Cox Regression based algorithms including stepwise Cox, Coxboost, plsRcox (Cox), Supervised Principal Components (SuperPC), Elastic Network (Enet), Gradient Boosting Machine (GBM).

Article	Cancer type								Machine learning algorithms															Datasets				Feature importance analysis		Machine learning task	
	EOC	ESC	NSC	LC	PC	CC	TC	BC	DT	KNN	SVM	NN	LASSO	RF	PCA-LDA	XGB	GLM	LR	Cox	SuperPC	Enet	TabNet	GBM	Self-produced	GEO	TCGA	EGA	Statistical	Feature ranking algorithms	Classification	Regression
Classification of paclitaxel-resistant ovarian cancer cells using holographic flow cytometry through interpretable machine learning [1]	+								+	+	+	+												+					+	+	
Heterogeneity of computational pathomic signature predicts drug resistance and intra-tumor heterogeneity of ovarian cancer [2]	+												+													+		+	+		+
Mitochondria-related chemoradiotherapy resistance genes-based machine learning model associated with immune cell infiltration on the prognosis of esophageal cancer and its value in pan-cancer [3]		+								+	+	+	+	+	+	+	+		+	+	+		+		+	+		+		+	+
Molecular separation-assisted label-free SERS combined with machine learning for nasopharyngeal cancer screening and radiotherapy resistance prediction [4]			+												+									+						+	
A Predictive Model for Initial Platinum-Based Chemotherapy Efficacy in Patients with Postoperative Epithelial Ovarian Cancer Using Tissue-Derived Small Extracellular Vesicles [5]	+												+					+							+	+			+	+	
Identifying genes associated with resistance to KRAS G12C inhibitors via machine learning methods [6]				+					+	+	+		+	+		+							+		+				+	+	
Turning to immunosuppressive tumors: Deciphering the immunosenescence-related microenvironment and prognostic characteristics in pancreatic cancer, in which GLUT1 contributes to gemcitabine resistance [7]					+						+		+						+	+	+		+		+	+	+	+			+
Machine learning-based models for genomic predicting neoadjuvant chemotherapeutic sensitivity in cervical cancer [8]						+								+										+				+		+	
A deep tabular data learning model predicting cisplatin sensitivity identifies BCL2L1 dependency in cancer [9]	+												+			+		+			+	+			+	+			+		+	
Deep neural networks identify signaling mechanisms of ErbB-family drug resistance from a continuous cell morphology space [10]		+		+			+	+				+		+	+									+						+	

Table 2. Results obtained in research papers. Abbreviations: Area under curve (AUC), Root Mean Squared Error (RMSE), Receiver Operating Characteristic (ROC), Matthew's correlation coefficient (MCC).

Article	Key results	Metrics					
		Accuracy	AUC	RMSE	F1	ROC	MCC
Classification of paclitaxel-resistant ovarian cancer cells using holographic flow cytometry through interpretable machine learning [1]	Demonstrated that morphological changes in epithelial ovarian cancer (EOC) cells correlate with drug sensitivity, highlighting the potential for monitoring drug resistance.	94.5%					
Heterogeneity of computational pathomic signature predicts drug resistance and intra-tumor heterogeneity of ovarian cancer [2]	Demonstrated a strong correlation between intra-tumor heterogeneity (ITH) and drug resistance in epithelial ovarian cancer (EOC) cells.		0.601				
Mitochondria-related chemoradiotherapy resistance genes-based machine learning model associated with immune cell infiltration on the prognosis of esophageal cancer and its value in pan-cancer [3]	Proposed a model that incorporates mitochondria-related chemoradiotherapy resistance (MRCRTR) genes. Identified six mitochondria-related genes that affect CRT and the prognosis of esophageal cancer.			0.001			
Molecular separation-assisted label-free SERS combined with machine learning for nasopharyngeal cancer screening and radiotherapy resistance prediction [4]	Developed a novel approach using label-free surface-enhanced Raman spectroscopy (SERS) to profile molecular patterns in the blood of nasopharyngeal cancer (NPC) patients, distinguishing those with radiotherapy sensitivity from those with resistance.	96.7%				0.999	
A Predictive Model for Initial Platinum-Based Chemotherapy Efficacy in Patients with Postoperative Epithelial Ovarian Cancer Using Tissue-Derived Small Extracellular Vesicles [5]	Found that three immune-related proteins—CCR1, IGHV3-35, and CD72—along with the presence of postoperative residual tumors, are strong predictors of platinum resistance in EOC patients. Proposed a model that can predict the efficacy of initial platinum-based chemotherapy.		0.960				
Identifying genes associated with resistance to KRAS G12C inhibitors via machine learning methods [6]	Identified some top-ranked genes, including H2AFZ, CKS1B, TUBA1B, RRM2, and BIRC5, associated with cancer progression and drug resistance. Have built efficient classifiers as the byproduct.	93.1%			0.938		0.860
Turning to immunosuppressive tumors: Deciphering the immunosenescence-related microenvironment and prognostic characteristics in pancreatic cancer, in which GLUT1 contributes to gemcitabine resistance [7]	Identified that IMSP1 and IMSP2 phenotypes influence pancreatic cancer prognosis and treatment response. Found that high MLIRS scores are linked to lower immune infiltration, while low scores indicate better drug sensitivity. Highlighted GLUT1 as a key factor driving tumor proliferation, migration, and chemotherapy resistance.		0.910				
Machine learning-based models for genomic predicting neoadjuvant chemotherapeutic sensitivity in cervical cancer [8]	Built random forest based classifier which can predict the response to platinum-based neoadjuvant chemotherapy. Through MDI feature analysis identified most significant SNPs — rs4558508, rs1130233, and rs725954. Found that they all are located within the Act gene family.						
A deep tabular data learning model predicting cisplatin sensitivity identifies BCL2L1 dependency in cancer [9]	Developed a deep learning model using the TabNet algorithm to predict cisplatin sensitivity based on cisplatin-perturbed gene expression data. Identified several key genes contributing to cisplatin resistance, most notably BCL2L1.	83.3%			0.831		0.51
Deep neural networks identify signaling mechanisms of ErbB-family drug resistance from a continuous cell morphology space [10]	Demonstated how deep learning of cell morphologies can be used to successfully predict drug resistance state in cancer cell lines from diverse tissues. Built a classifier based on deep neural network and random forest which can identify cancer cells resistance to ErbB-family drugs.	74%					

References

[1] L. Xin et al., "Classification of Paclitaxel-resistant Ovarian Cancer Cells Using Holographic Flow Cytometry through Interpretable Machine Learning," Sensors and Actuators B Chemical, vol. 414, p. 135948, May 2024, doi: 10.1016/j.snb.2024.135948.

[2] Q. Zhu et al., "Heterogeneity of computational pathomic signature predicts drug resistance and intra-tumor heterogeneity of ovarian cancer," Translational Oncology, vol. 40, p. 101855, Jan. 2024, doi: 10.1016/j.tranon.2023.101855.

[3] Z. Liu et al., "Mitochondria-related chemoradiotherapy resistance genes-based machine learning model associated with immune cell infiltration on the prognosis of esophageal cancer and its value in pan-cancer," Translational Oncology, vol. 42, p. 101896, Feb. 2024, doi: 10.1016/j.tranon.2024.101896.

[4] J. Zhang et al., "Molecular separation-assisted label-free SERS combined with machine learning for nasopharyngeal cancer screening and radiotherapy resistance prediction," Journal of Photochemistry and Photobiology B Biology, vol. 257, p. 112968, Jun. 2024, doi: 10.1016/j.jphotobiol.2024.112968.

[5] S. Shen et al., "A Predictive Model for Initial Platinum-Based Chemotherapy Efficacy in Patients with Postoperative Epithelial Ovarian Cancer Using Tissue-Derived Small Extracellular Vesicles," Journal of Extracellular Vesicles, vol. 13, no. 8, Aug. 2024, doi: 10.1002/jev2.12486.

[6] X. Lin et al., "Identifying genes associated with resistance to KRAS G12C inhibitors via machine learning methods," Biochimica Et Biophysica Acta (BBA) - General Subjects, vol. 1867, no. 12, p. 130484, Oct. 2023, doi: 10.1016/j.bbagen.2023.130484.

[7] S.-Y. Lu et al., "Turning to immunosuppressive tumors: Deciphering the immunosenescence-related microenvironment and prognostic characteristics in pancreatic cancer, in which GLUT1 contributes to gemcitabine resistance," Heliyon, vol. 10, no. 17, p. e36684, Aug. 2024, doi: 10.1016/j.heliyon.2024.e36684.

[8] L. Guo, W. Wang, X. Xie, S. Wang, and Y. Zhang, "Machine learning-based models for genomic predicting neoadjuvant chemotherapeutic sensitivity in cervical cancer," Biomedicine & Pharmacotherapy, vol. 159, p. 114256, Jan. 2023, doi: 10.1016/j.biopha.2023.114256.

[9] A. Nasimian, M. Ahmed, I. Hedenfalk, and J. U. Kazi, "A deep tabular data learning model predicting cisplatin sensitivity identifies BCL2L1 dependency in cancer," Computational and Structural Biotechnology Journal, vol. 21, pp. 956–964, doi: 10.1016/j.csbj.2023.01.020.

[10] J. Longden et al., "Deep neural networks identify signaling mechanisms of ErbB-family drug resistance from a continuous cell morphology space," Cell Reports, vol. 34, no. 3, p. 108657, Jan. 2021, doi: 10.1016/j.celrep.2020.108657.

[11] International Agency for Research on Cancer, F. Bray, IARC, E. Weiderpass, and World Health Organization, "Latest global cancer data: Cancer burden rises to 19.3 million new cases and 10.0 million cancer deaths in 2020," IARC, Dec. 15, 2020. https://www.iarc.who.int/wp-content/uploads/2020/12/pr292_E.pdf (accessed Dec. 01, 2024).

[12] Sh. Huang and B. O. Sullivan, "Oral cancer: Current role of radiotherapy and chemotherapy," Medicina Oral, Patología Oral Y Cirugía Bucal, pp. e233–e240, Jan. 2013, doi: 10.4317/medoral.18772.

[13] L. Kuroki and S. R. Guntupalli, "Treatment of epithelial ovarian cancer," BMJ, p. m3773, Nov. 2020, doi: 10.1136/bmj.m3773.

[14] S. W. Johnson, R. F. Ozols, and T. C. Hamilton, "Mechanisms of drug resistance in ovarian cancer," Cancer, vol. 71, no. S2, pp. 644–649, Aug. 2010, doi: 10.1002/cncr.2820710224.

[15] Y. Jiang, M. Yang, S. Wang, X. Li, and Y. Sun, "Emerging role of deep learning-based artificial intelligence in tumor pathology," Cancer Communications, vol. 40, no. 4, pp. 154–166, Apr. 2020, doi: 10.1002/cac2.12012.

[16] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," arXiv (Cornell University), Jan. 2017, doi: 10.48550/arxiv.1705.07874.

[17] R. Tibshirani, "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society Series B (Statistical Methodology), vol. 58, no. 1, pp. 267–288, Jan. 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.

[18] C. McQuin et al., "CellProfiler 3.0: Next-generation image processing for biology," PLoS Biology, vol. 16, no. 7, p. e2005970, Jul. 2018, doi: 10.1371/journal.pbio.2005970.

[19] "The Cancer Genome Atlas Program (TCGA)," Cancer.gov. https://www.cancer.gov/ccg/research/genome-sequencing/tcga (accessed Dec. 01, 2024).

[20] "Gene Expression Omnibus (GEO) Database." https://www.ncbi.nlm.nih.gov/geo/ (accessed Dec. 01, 2024).

[21] "EGA European Genome-Phenome Archive," The European Bioinformatics Institute (EMBL-EBI). https://ega-archive.org/ (accessed Dec. 01, 2024).

[22] "ATCC: The Global Bioresource Center," ATCC. https://www.atcc.org/ (accessed Dec. 01, 2024).

[23] "R: The R Project for Statistical Computing." https://www.r-project.org/ (accessed Dec. 01, 2024).

[24] P. Biecek, "DALEX: Explainers for Complex Predictive Models in R," Zenodo (CERN European Organization for Nuclear Research), Feb. 2020, doi: 10.5281/zenodo.3670940.

[25] "PerkinElmer | Science with purpose." https://content.perkinelmer.com/ (accessed Dec. 01, 2024).