

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

# Introduction

Progress has been made in chemotherapy drugs, but drug resistance remains a major challenge in cancer treatment and the main cause of cancer progression and even death. However, there are no clear indicators for predicting the risk of drug resistance in patients. Existing drug sensitivity assessment methods has limitations such as low modeling success rates, high cost, and time-consuming process. Machine learning is both an expanding and evolving field of computing, and it seems that it can significantly help in solving chemotherapy resistance problem. Here we provide an overview of how different studies apply machine learning algorithms to predict and understand chemotherapy resistance in various cancer types. Also we consider the strengths and limitations of each approach and discuss obtained results.

# 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 similar 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 [8] 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 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:

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

where:  $\text{expr}_{\text{gene}}$  denotes the expression level of each gene,  $\text{coeff}_{\text{gene}}$  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.

## 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 [10] and The Cancer Genome Atlas (TCGA) database [9], which provide researchers with access to extensive datasets. Moreover, nonprofit organizations like the American Type Culture Collection (ATCC) [12] enable researchers to obtain biological materials, including cancer cells.

Authors of [1] prepared their dataset specifically for their research. Four kinds of epithelial ovarian cancer cells with different drug sensitivity (SKOV3, SKOV3\_Ta\_2 $\mu$ M, SKOV3\_Ta\_8 $\mu$ M, and SKOV3\_Ta\_20 $\mu$ M) were studied in this work. The SKOV3 cells were sourced from the ATCC [12] and preserved at the Obstetrics and Gynecology Laboratory of Peking University People's Hospital. The drug-resistant characteristics of SKOV3\_Ta\_2 $\mu$ M, SKOV3\_Ta\_8 $\mu$ M, and SKOV3\_Ta\_20 $\mu$ M 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 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.

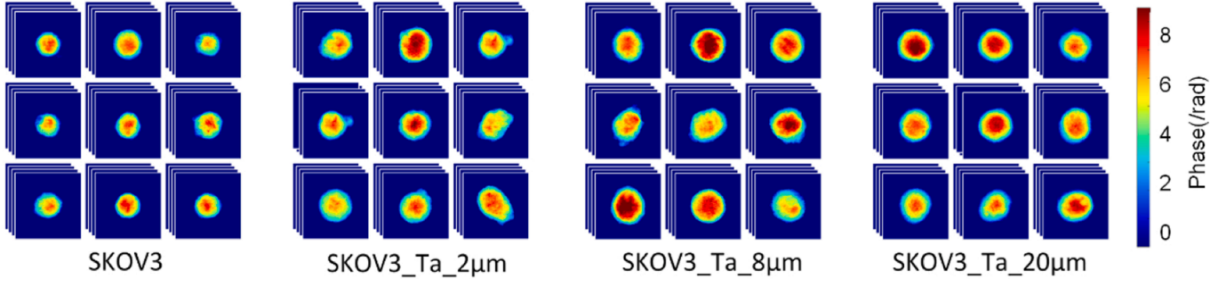


Figure 1. Reconstructed QPIs of EOC cells used by authors of [1].

In article [4], same as in [1], authors chose the approach of collecting their own dataset. Their dataset 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 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.

Authors of articles [2], [3], [6] and [7] 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 [9]. 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 [10]. 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 [11]. In this study they used

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 [10]. 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 [9], was used.

Table 1. Methods used in research papers.

Article	Cancer type	Machine learning algorithms	Datasets	Feature importance analysis
Classification of paclitaxel-resistant ovarian cancer cells using holographic flow cytometry through interpretable machine learning [1]	Epithelial ovarian cancer (EOC)	Tree, Naive Bayes, K-nearest neighbors (KNN), support vector machine (SVM), and neural network (NN)	Self-produced dataset of 2998 quantitative phase images (QPIs) of EOC cells	SHapley Additive exPlanations (SHAP), Pearson coefficient, Kruskal-Wallis test
Heterogeneity of computational pathomic signature predicts drug resistance and intra-tumor heterogeneity of ovarian cancer [2]	Epithelial ovarian cancer (EOC)	CellProfiler [8], least absolute shrinkage and selection operator (LASSO) regression	494 ovarian and 70 paracarcinoma tissues images from The Cancer Genome Atlas (TCGA) database [9]	Statistical analysis using R [13]. Various visualizations, including heatmaps, Venn diagrams, ROC curves, and survival curves.
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]	Esophageal cancer	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)	Nearly 500 tissue samples, RNA-sequences and some other clinical data from Gene Expression Omnibus (GEO) database [10], information on 183 esophageal cancer patients from The Cancer Genome Atlas (TCGA) database [9]	Statistical analysis using DALEX package [14] for R [13]
A Predictive Model for Initial Platinum-Based Chemotherapy Efficacy in Patients with Postoperative Epithelial Ovarian Cancer Using Tissue-Derived Small Extracellular Vesicles [5]	Epithelial ovarian cancer (EOC)	Least absolute shrinkage and selection operator (LASSO) regression, logistic regression (LR)	Nearly 300 tissue samples, and other clinical data from Gene Expression Omnibus (GEO) database [10], transcriptomic sequencing data and clinical information from tumor tissues of 379 EOC patients from The Cancer Genome Atlas (TCGA) database [9]	

Table 1. Methods used in research papers (continued).

Article	Cancer type	Machine learning algorithms	Datasets	Feature importance analysis
Molecular separation-assisted label-free SERS combined with machine learning for nasopharyngeal cancer screening and radiotherapy resistance prediction [4]	Nasopharyngeal cancer	Principal component analysis and linear discriminant analysis (PCA-LDA)	Self-produced dataset of 120 plasma samples, 60 of which from healthy volunteers, 30 from radiotherapy sensitivity patients and 30 from radiotherapy resistance patients	
Identifying genes associated with resistance to KRAS G12C inhibitors via machine learning methods [6]	Lung cancer	Random forest (RF), support vector machine (SVM), K-nearest neighbors (KNN), decision tree (DT)	7612 sample of gene expression profile data from Gene Expression Omnibus (GEO) database [10]. Each sample contained the expression of 8687 genes	Seven feature ranking algorithms were applied, 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 (CATB), and extreme gradient boosting (XGB)
Turning to immunosuppressive tumors: Deciphering the immunosenescence-related microenvironment and prognostic characteristics in pancreatic cancer, in which GLUT1 contributes to gemcitabine resistance [7]	Pancreatic cancer	Support vector machine (SVM), CoxBoost, random forest (RF), least absolute shrinkage and selection operator (LASSO), stepwise Cox, partial least squares regression for Cox (plsRcox), Ridge, supervised principal components (SuperPC), elastic network (Enet), generalized boosted regression modeling (GBM)	Nearly 1000 samples from 12 datasets from The Cancer Genome Atlas (TCGA) [9], Gene Expression Omnibus (GEO) [10] and The European Genome-phenome Archive (EGA) [11]	The univariate Cox regression analysis was used to identify immunosenescence-related genes with prognostic significance in pancreatic cancer. Genes with a p-value of less than 0.01 were selected as meaningful features for subsequent analysis



Table 2. Machine learning algorithms comparison. Algorithms: 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	DT	KNN	SVM	NN	LASSO	RF	PCA-LDA	XGB	GLM	LR	Cox	SuperPC	Enet	GBM
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]			+		+						+	+	+	+

Table 3. Results obtained in research papers.

Article	Key results	Best algorithms	Metrics
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.	Support vector machine (SVM) and neural network (NN)	Accuracy of 94.5% for SVM and 93.4% for NN
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	Least absolute shrinkage and selection operator (LASSO) regression	Area under curve (AUC) of 0.601, 0.594, and 0.589 for 1, 3, and 5 years survival time accordingly
Mitochondria-related chemoradiotherapy resistance genes-based machine learning model associated with immune cell infiltration on the prognosis of esophageal cancer and its value in pancreatic 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.	Neural network (NN) and least absolute shrinkage and selection operator (LASSO) regression	Root mean squared error (RMSE) of 0.001 for NN and 0.003 for LASSO
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	Principal component analysis and linear discriminant analysis (PCA-LDA)	Accuracy of 96.7% for identifying radiotherapy resistance subjects from sensitivity ones and 100% for identifying the nasopharyngeal cancer (NPC) subjects from healthy ones

Table 3. Results obtained in research papers (continued).

Article	Key results	Best algorithms	Metrics
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	Least absolute shrinkage and selection operator (LASSO) regression and logistic regression (LR)	Area under curve (AUC) of 0.864
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	Categorical boosting (CATB) for feature selection and support vector machine (SVM) for classification	Accuracy of 93.1% and F1-score of 0.938
Turning to immuno-suppressive 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	Stepwise Cox combined with generalized boosted regression modeling (GBM)	Area under the curve (AUC) of 0.91

# References

- [1] Lu Xin, Wen Xiao, Huanzhi Zhang, Yakun Liu, Xiaoping Li, Pietro Ferraro, Feng Pan, Classification of paclitaxel-resistant ovarian cancer cells using holographic flow cytometry through interpretable machine learning, 2024.
- [2] Qiuli Zhu, Hua Dai, Feng Qiu, Weiming Lou, Xin Wang, Libin Deng, Chao Shi, Heterogeneity of computational pathomic signature predicts drug resistance and intra-tumor heterogeneity of ovarian cancer, 2024.
- [3] Ziyu Liu, Zahra Zeinalzadeh, Tao Huang, Yingying Han, Lushan Peng, Dan Wang, Zongjiang Zhou, DIABATE Ousmane, Junpu Wang, 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, 2024.
- [4] Jun Zhang, Youliang Weng, Yi Liu, Nan Wang, Shangyuan Feng, Sufang Qiu, Duo Lin, Molecular separation-assisted label-free SERS combined with machine learning for nasopharyngeal cancer screening and radiotherapy resistance prediction, 2024.
- [5] Shen S, Wang C, Gu J, Song F, Wu X, Qian F, Chen X, Wang L, Peng Q, Xing Z, Gu L, Wang F, Cheng X. A Predictive Model for Initial Platinum-Based Chemotherapy Efficacy in Patients with Postoperative Epithelial Ovarian Cancer Using Tissue-Derived Small Extracellular Vesicles, 2024.
- [6] Xiandong Lin, QingLan Ma, Lei Chen, Wei Guo, Zhiyi Huang, Tao Huang, Yu-Dong Cai, Identifying genes associated with resistance to KRAS G12C inhibitors via machine learning methods, 2023.
- [7] Si-Yuan Lu, Qiong-Cong Xu, De-Liang Fang, Yin-Hao Shi, Ying-Qin Zhu, Zhi-De Liu, Ming-Jian Ma, Jing-Yuan Ye, Xiao Yu Yin, Turning to immunosuppressive tumors: Deciphering the immunosenescence-related microenvironment and prognostic characteristics in pancreatic cancer, in which GLUT1 contributes to gemcitabine resistance, 2024.
- [8] T. Misteli, C. McQuin, A. Goodman, V. Chernyshev, L. Kametsky, B.A. Cimini, et al., CellProfiler 3.0: next-generation image processing for biology, 2018.
- [9] The Cancer Genome Atlas (TCGA) database. Available at <https://www.cancer.gov/ccg/research/genome-sequencing/tcga>. Accessed October 8, 2024.
- [10] Gene Expression Omnibus (GEO) database. Available at <https://www.ncbi.nlm.nih.gov/geo/>. Accessed October 8, 2024.
- [11] The European Genome-phenome Archive (EGA). Available at <https://ega-archive.org/>. Accessed October 8, 2024.
- [12] American Type Culture Collection (ATCC). Available at <https://www.atcc.org/>. Accessed October 8, 2024.
- [13] The R Project for Statistical Computing. Available at <https://www.r-project.org/>. Accessed October 8, 2024.
- [14] DALEX: explainers for complex predictive models, Przemyslaw Biecek, 2018.