A Bioinformatics Analysis of Ovarian Cancer Data Using Machine Learning

Vincent Schilling *, Peter Beyerlein , Jeremy Chien

Posted Date: 6 May 2023

doi: 10.20944/preprints202305.0413.v1

Keywords: ovarian cancer; machine learning; SHAP; diagnostic biomarkers; platinum resistance

Article

A Bioinformatics Analysis of Ovarian Cancer Data Using Machine Learning

Vincent Schilling [1,3,*], Peter Beyerlein [2] and Jeremy Chien [3]

[1] Technical University of Applied Sciences Wildau; vschilling@ucdavis.edu
[2] ibiomics UG, peter.beyerlein@googlemail.com
[3] Department of Biochemistry and Molecular Medicine, University of California Davis; jrchien@ucdavis.edu
* Correspondence: vschilling@ucdavis.edu

Abstract: The identification of biomarkers is crucial for cancer diagnosis, understanding the underlying biological mechanisms, and developing targeted therapies. In this study we propose a machine learning approach to predict the outcome and platinum resistance status of ovarian cancer patients using public available gene expression data. Six classical machine learning algorithms are compared on their predictive performance. Those with the highest score are analyzed by their feature importance using the SHAP algorithm. We were able to select multiple genes that were correlating with the outcome and platinum resistance status of the patients and validated those using Kaplan-Meier plots. In comparison to similar approaches the performance of the models were higher and different genes using feature importance analysis were identified. The most promising identified genes that could be used as biomarkers are: TMEFF2, ACSM3, SLC4A1 and ALDH4A1.

Keywords: ovarian cancer; machine learning; SHAP; diagnostic biomarkers; platinum resistance

1. Introduction

Ovarian Cancer is the most fatal gynecologic malignancy with a five year survival rate from the year 2011-2017 of 49% considering all stages of the cancer [1]. The cancer is very aggressive and often recurs after subsequent treatment for these recurrences. Most patients will acquire resistance through treatment consisting of carboplatin based chemotherapy as well as PARP inhibitors [2,3].

Another property of the cancer which leads to such poor survival rates is the comparatively late detection of it. Most of the time when patients get diagnosed they are already in the advanced stages III and IV. The symptoms are very vague and it is hard to identify for medical professionals if these are indication of ovarian cancer. Therefore early detection methods, genetic screening and multiple treatment options are very important in the fight and treatment of this cancer [4,5].

In this paper the biggest publicly available data set of ovarian cancer with over 585 patients is analyzed to predict the outcome of ovarian cancer patients based on their gene expression value and find possible targets for targeted therapy and biomarkers. The same dataset is used to predict the platinum resistance status of ovarian cancer patients to find biomarkers as well. A combination of bioinformatics analysis and machine learning methods is used to identify relationships between biological components and the progression of the patients.

In recent studies it has already been demonstrated that biological parameters like mRNA gene expression can be linked to and predict the outcome of cancer patients [6–8]. For that matter statistical methods have been used as well as machine learning methods [9].

The advantage of using computational methods is that they are faster than conventional methods like shutting down genes and evaluate it on living cells if they have an effect on the fitness of the cancer. Because even though computational methods might not give an exact answer whether a biological component can be used as a biomarker or not it can limit the amount of potential candidates and can identify relationships between the biological components and give more conclusive answers [10].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2023 doi:10.20944/preprints202305.0413.v1

The idea of evaluating machine learning models to identify which features have the biggest impact on the decision is still pretty new but scientist in the bioinformatics field started to use it to identify important biological features in big datasets [11–13].

The first approach is to identify patients in the dataset that had a very poor progression with those that had very good progression of the disease to have a very distinct comparison of the biology between those. To identify the targets that have a very high impact on the outcome of the patients a method from the area of explainable artificial intelligence will be used to analyze the machine learning models and check what inputs have the biggest weight on the outcome of the patients to potentially find biomarkers. The second approach is to execute the same procedure on data from patients that have been classified as either platinum sensitive or platinum resistance. The performance of the different machine learning methods will be compared.

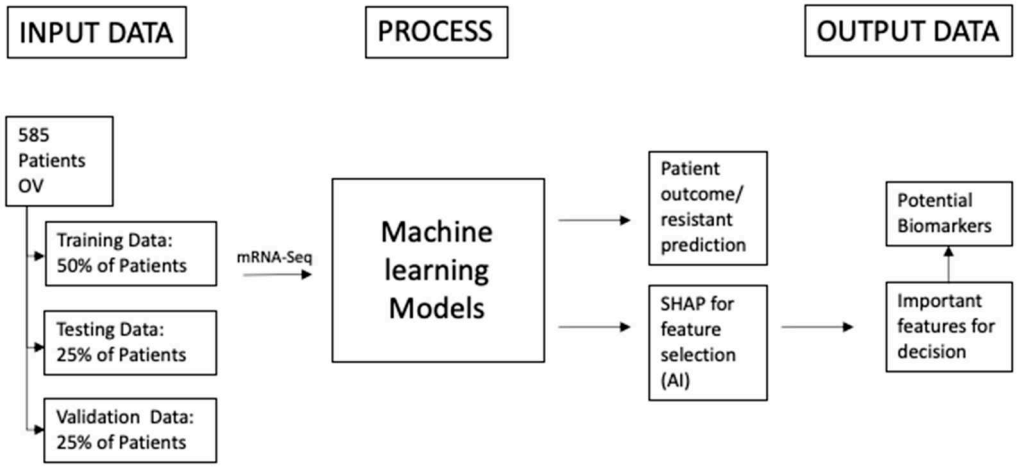


Figure 1. The schematic shows the overall project structure. The study on ovarian cancer consists overall of 585 patients but not for every patient all biological and clinical data is available. The main idea is that the patients will be put into two very distinct categories for example: good outcome, bad outcome or resistant and sensitive. After putting the patients into two classes they will be distributed into three datasets: 50% Training Data, 25% Testing Data and 25% Validation Data. With that method there won't be underlying fitting to the data when the models are getting adjusted and tested. The performance of the models will be determined with a one-time test on the validation data set. The patients and their features are the input for the machine learning models. They should predict the class of the patients. After that when the performance of the models is sufficient algorithms from the field of explainable artificial intelligence will be applied to detect which biological features contribute the most to the decision of the model. Those features will most likely have a biological implication why the patient is having a good or bad outcome.

2. Materials and Methods

2.1. Materials

The data used in the analysis is from the cancer genome atlas (TCGA) [14,15]. It consists of 585 ovarian cancer patients. The unnormalized gene counts for the mRNA sequencing data from GDAC is used for differential gene expression analysis with DESeq2 [16]. Apart from the raw gene counts the clinical data of the patients has been collected to determine whether the patients had a good or bad outcome and if they developed a platinum resistance or were sensitive for it.

Due to limitations in available data, the outcome determination and platinum resistance status were not available for all patients in this study. To address this, two approaches were taken. For the first approach, patients were filtered based on the availability of mRNA sequencing data and their classification as either having a good or bad outcome. In the second approach, patients were required to have both mRNA sequencing data and platinum resistance status available. After filtering, the remaining patient data was randomly divided into training, testing, and independent validation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2023 doi:10.20944/preprints202305.0413.v1

datasets. The training dataset was utilized to train the machine learning models, whereas the testing dataset was used to test the performance of these models in multiple trials. Finally, the validation dataset was employed to evaluate the performance of the models on data that was not previously used for training or testing.

2.1.1. Data on the outcome of the patients:

To identify potential biomarkers for ovarian cancer, clear differentiation between patients with good and bad outcome following treatment is essential. As such, patients were classified as having a bad outcome if they had died within two years after treatment, while those who survived for five years or more were classified as having a good outcome. These specific timeframes were selected in order to facilitate clear separation between the patient groups, enabling more distinct differential gene expression analysis. The study cohort included a total of 113 patients, of which 55 were classified as having a good outcome and 58 as having a bad outcome.

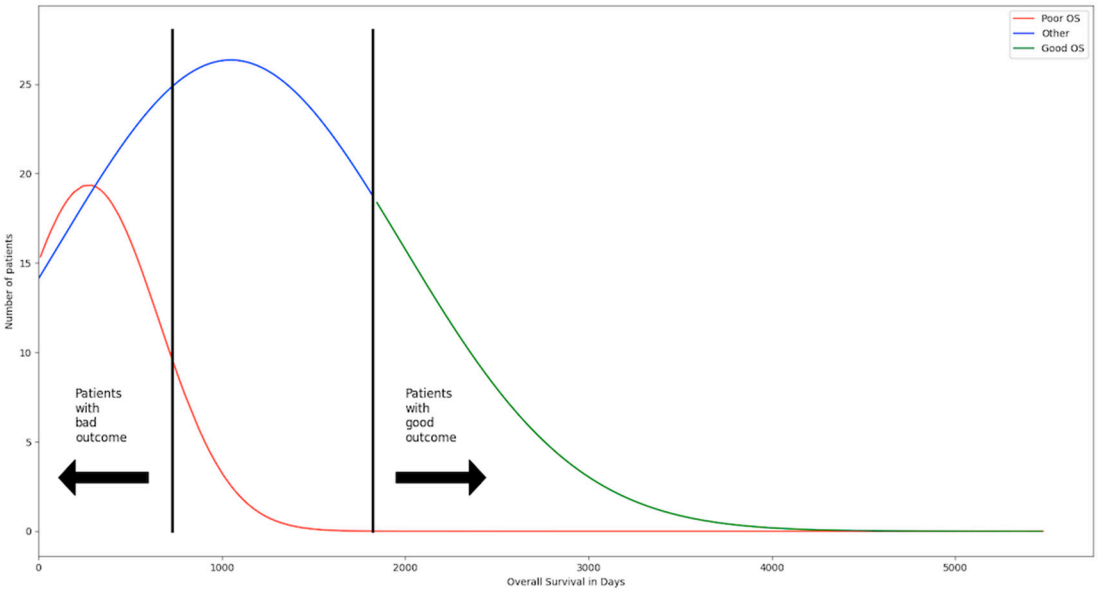


Figure 2. The patients have been separated in three groups. The first group are the patients with poor overall survival (OS) and consists of 112 patients. They died within the first two years after treatment and are considered to have a bad outcome. The second group are the patients with positive OS and consists of 119 patients. They have lived five years and longer after treatment and are considered to have a good outcome. The third group are all the patients in between the two groups and make up the biggest portion with 345. Those patients cannot be considered good or bad outcome. The threshold has been selected to have two groups that are very distinct from one another to make changes and differences in their gene expression profile easier to detect.

2.1.2. Data on the resistance status of the patients:

To identify genes that could predict the platinum sensitivity in ovarian cancer the categorization from the TCGA has been used. In total there are 152 patients. 109 of those are sensitive and 43 are resistant.

2.2. Methods

2.2.1. Machine learning methods:

The machine learning methods have been all trained with 50% of the data from each dataset and then frequently tested on 25% of the remaining ones. When the performance reached the desired value it has been tested ones on the remaining 25% to evaluate how well the method performs on unseen data. All of the methods have been trained as binary classifier to identify the correct class by

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2023 doi:10.20944/preprints202305.0413.v1

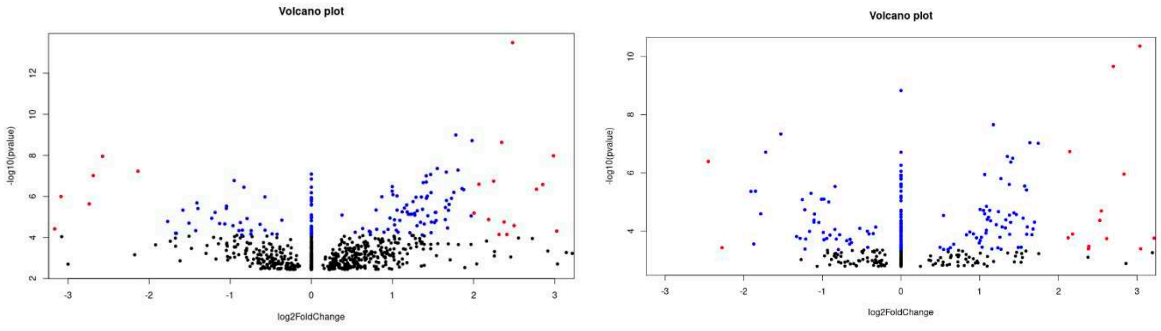
their gene expression profile. The models have been tested with the normalized data and data from the principal component analysis. The training of the models has been done in different Jupyter Notebooks on the "cancer genomics cloud". For each machine learning model the f1-score has been calculated to determine their predictive performance. In the result part the confusion matrix of the classifiers are depicted to see how their classification performance is distributed via the two classes. The machine learning methods that have been used are: K-means clustering [17,18], Naïve Bayes [19], logistic regression [20–22], supported vector machines [23,24], Random Forest [25,26] and XBGoost [27]. All the used methods were integrated via the sklearn library in python. Different hyperparameters have been used to adjust the model to the gene expression data.

2.2.2. SHAP

To determine which genes can be used as biomarkers the XAI method SHAP is used. It utilizes the mathematics from game theory to determine the global significance of the genes for the decision making of the model. After the model has been trained each gene will be given a Shapley value the determines their significance and depending on their overall expression between the two classes it can be assessed if the expression of the gene is correlated to the outcome or platinum resistant status of a patient [28].

2.2.3. Bioinformatics algorithms

The original gene expression data consists of 20429 genes. Most of them are likely not relevant in the search of biomarkers because they will have no biological function associated with the fitness of the cancer. Those genes will not be differentially expressed between the two classes. DESeq2 is a software available in R used to determine differentially expressed genes between two groups. Unnormalized gene counts are used for that purpose. Afterwards the log2fold shrinkage method apeglm is used to determine which genes have significant changes between the two groups [29]. The p-adjusted value is used to select the genes that are significantly expressed between the two groups. A p-adjusted value of 0.01 for the outcome group has been set and 0.05 for the platinum resistant status group. For the outcome group, genes upregulated in patients with a good outcome have a log2foldchange above zero and the ones upregulated in patients with bad outcome have a log2foldchange lower than zero. For the platinum resistant status group, genes upregulated in patients that are platinum sensitive have a log2foldchange above zero and the ones that are upregulated in patients that are platinum resistant have a log2foldchange lower than zero. The selected gene lists will be further analyzed to determine their biological implications in relation to ovarian cancer.



(**a**) (**b**)

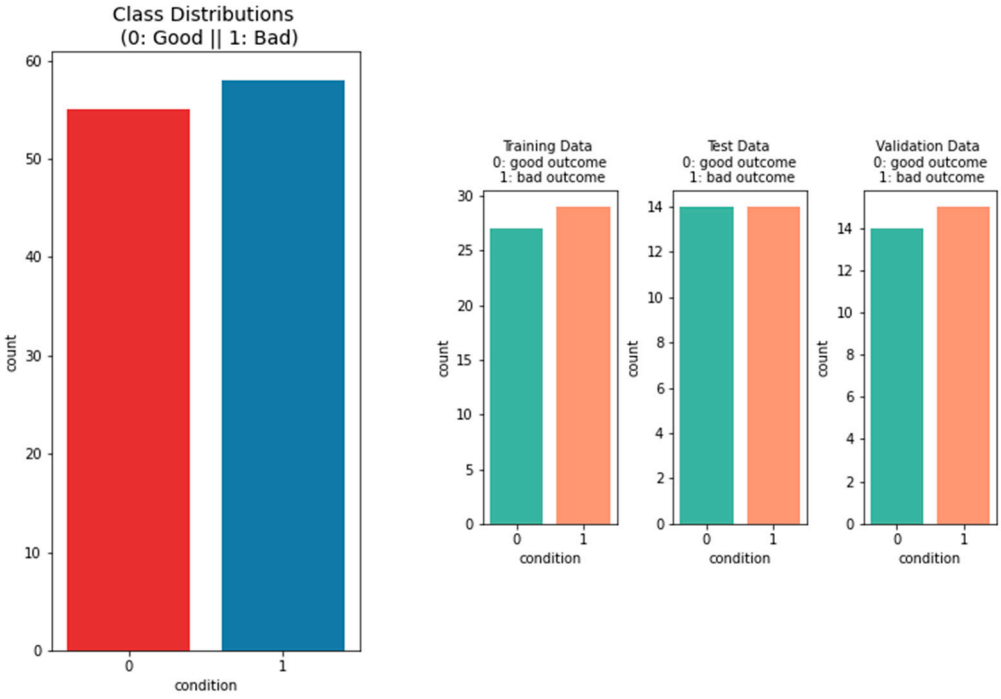
Figure 3. (a) The volcano plot depicted shows the differentially expressed genes between the patients with bad outcome and good outcome. The dots in blue represent genes that have a adjusted p-value of 0.01> and a log2FoldChange between +2 and –2. It means that they are significantly differentially expressed. The red dots represent genes that have a adjusted p-value of 0.01> and a log2FoldChang >2 and –2>. It means that they are highly significantly expressed. The black dots are all genes the have a higher adjusted p-value than 0.1. (b) The volcano plot depicted shows the differentially expressed genes between the patients that are sensitive and resistant. The dots in blue represent genes that have

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2023 doi:10.20944/preprints202305.0413.v1

a adjusted p-value of 0.05> and a log2FoldChange between +2 and –2. It means that they are significantly differentially expressed. The red dots represent genes that have a adjusted p-value of 0.05> and a log2FoldChang >2 and –2>. It means that they are highly significantly expressed. The black dots are all genes the have a higher adjusted p-value than 0.1.

2.2.4. Statistical methods

After the differential gene expression analysis the gene count data is normalized before it can be used by the machine learning methods. Since the data from the ovarian cancer patients has been collected by multiple hospitals it is need the normalize the sample within the gene and the sample itself. The methods for normalization that can be used is either the commonly used z-normalization or the TMM normalization [30]. Both methods have been used and the performance of the machine learning models was better using the TMM normalization method. Even after normalization there were still outliers so statistical method winsorizing has been applied to the data. So the highest 2.5% of the data is replaced with value right below them. After that SMOTE has been used on the training dataset of the platinum resistance status group [31]. Since the dataset is very imbalanced between the patients that are resistant and the ones that are sensitive the machine learning methods will not be performing well since the prediction will be more based on the class that is more frequent instead of the gene expression values. Therefore SMOTE is used to create fake data points between the original data points. With this approach the machine learning models will be trained with similar data as the original data points and won't lean to majority class while the integrity of the data stays intact. Afterwards all data points for both groups and datasets will be scaled between zero and one because some methods like SVM will perform better with that scaling. On the resulting datasets a PCA is performed since Random Forest and XGBoost classifiers worked better in the approach than without it. This is essentially just for the assessment of the best performer since it is very difficult to transform the principal components later back to the original features [32,33].

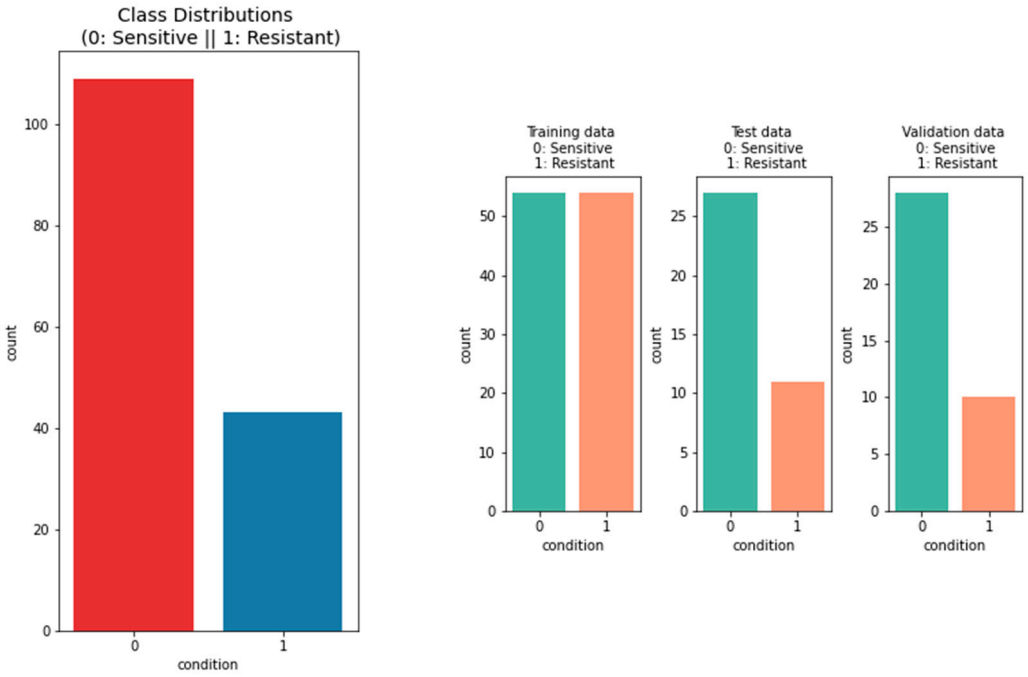


(a) **(b)**

Figure 4. The left bar plot depicts the overall distribution on the data . There are 55 patients that are considered to have good outcome and 57 patients that are considered to have a bad outcome. The distribution between the two classes is very even, so there is no need for over- or under-sampling. In

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2023 doi:10.20944/preprints202305.0413.v1

the bar plots on the right the distribution of the classes is shown between the training data, test data and validation data.



(a) **(b)**

Figure 5. The left bar plot depicts the overall distribution on the data . There are 109 patients that are considered to be sensitive and 43 patients that are considered to be resistant. The distribution between the two classes is uneven. Oversampling is used to adjust that in the trainings data. In the bar plots on the right the distribution of the classes is shown between the training data, test data and validation data. .

3. Results

3.1. Outcome prediction of ovarian cancer patients

3.1.1. Machine learning models performance:

Six different machine learning methods were used to predict the outcome of ovarian cancer patients. The performance is evaluated by the f1-score of the model on the validation dataset. The best performing model is the logistic regression model with 4 misclassifications out of 29 observations. The performance of the other models were slightly lower. The worst performing model is the K-means-clustering model. For the selection of the genes the models have been tested on the differentially expressed genes with a p-adjusted value of 0.1, 0.05 and 0.01. The gene dataset with a cutoff of 0.01 consisting of 149 genes was the best performing one.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2023 doi:10.20944/preprints202305.0413.v1

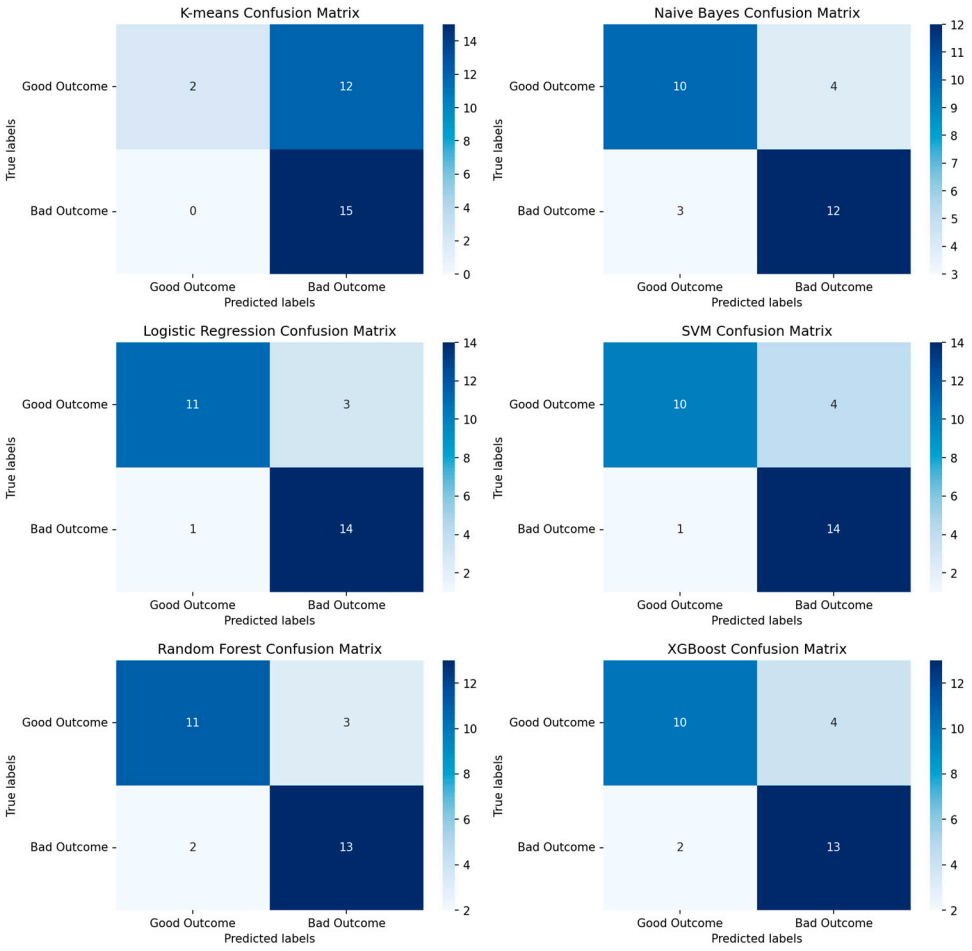


Figure 6. Depicted confusion matrixes for the different machine learning methods which have been used. The higher the number for the corresponds between true labels and predicted labels the more accurate and trustworthy the model is. Following are the f1-scores that evaluates the performance of the model: K-means Clustering: 0.71, Naïve Bayes: 0.77, Logistic Regression: 0.88, Supported Vector Machine: 0.85, Random Forest: 0.84, XGBoost: 0.82. The logistic regression model is the model with the highest f1-score and 4 misclassifications from 29 observations.

Table 1. Overview of the performance of machine learning models for the outcome prediction of ovarian cancer patients.

Machine learning Method	f1-score
K-means clustering	0.71
Naive Bayes	0.77
SVM	0.85
Logistic Regression	0.88
Random Forest	0.84
XGBoost	0.82

3.1.2. Identified genes

The SHAP Algorithm has been performed on the logistic regression model to evaluate which genes have the highest impact on the outcome of the model. The genes with the highest impact on the model are depicted in Figure 7A. The top 20 genes with the highest Shapley values can be found in Figure A1. Those genes were then separated into the ones that are upregulated in the good outcome group and those that are upregulated in the bad outcome group. Upregulated in this context means

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2023 doi:10.20944/preprints202305.0413.v1

the overall expression of that gene is higher in one group of patients than in the other. In Figure 7B the Kaplan-Meier plots are depicted from six genes with high Shapley values and low adjusted p-values [34]. The first two are upregulated in the group of patients with bad outcome and the other four are upregulated in the group of patients with good outcome.

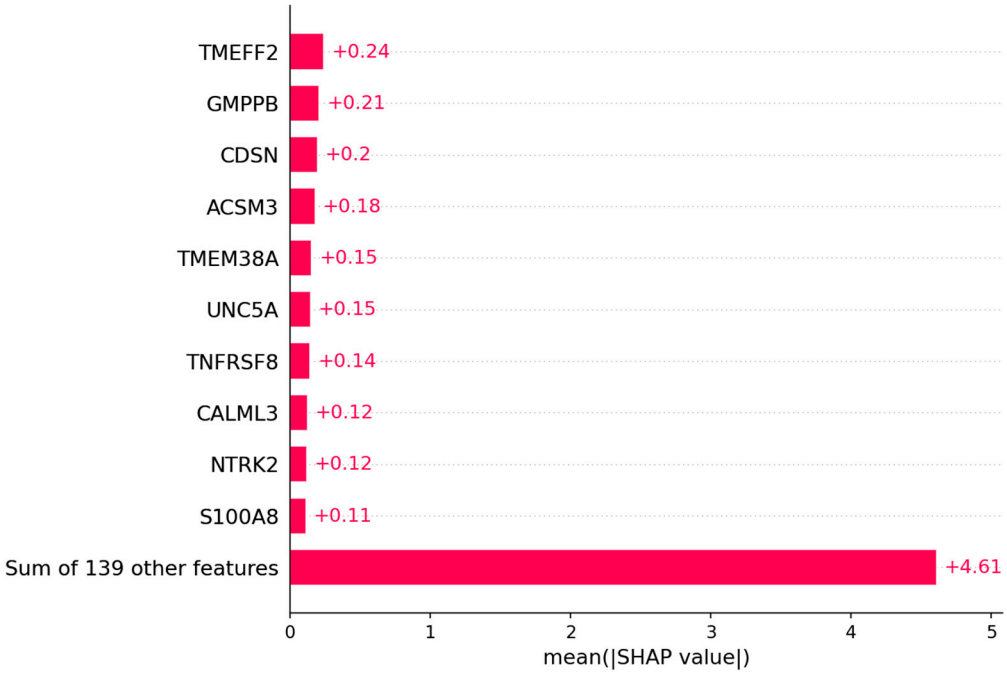


Figure 7. The bar plot shows the top 10 genes based on the mean SHAP value of the logistic regression model of the ovarian cancer outcome prediction.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2023 doi:10.20944/preprints202305.0413.v1

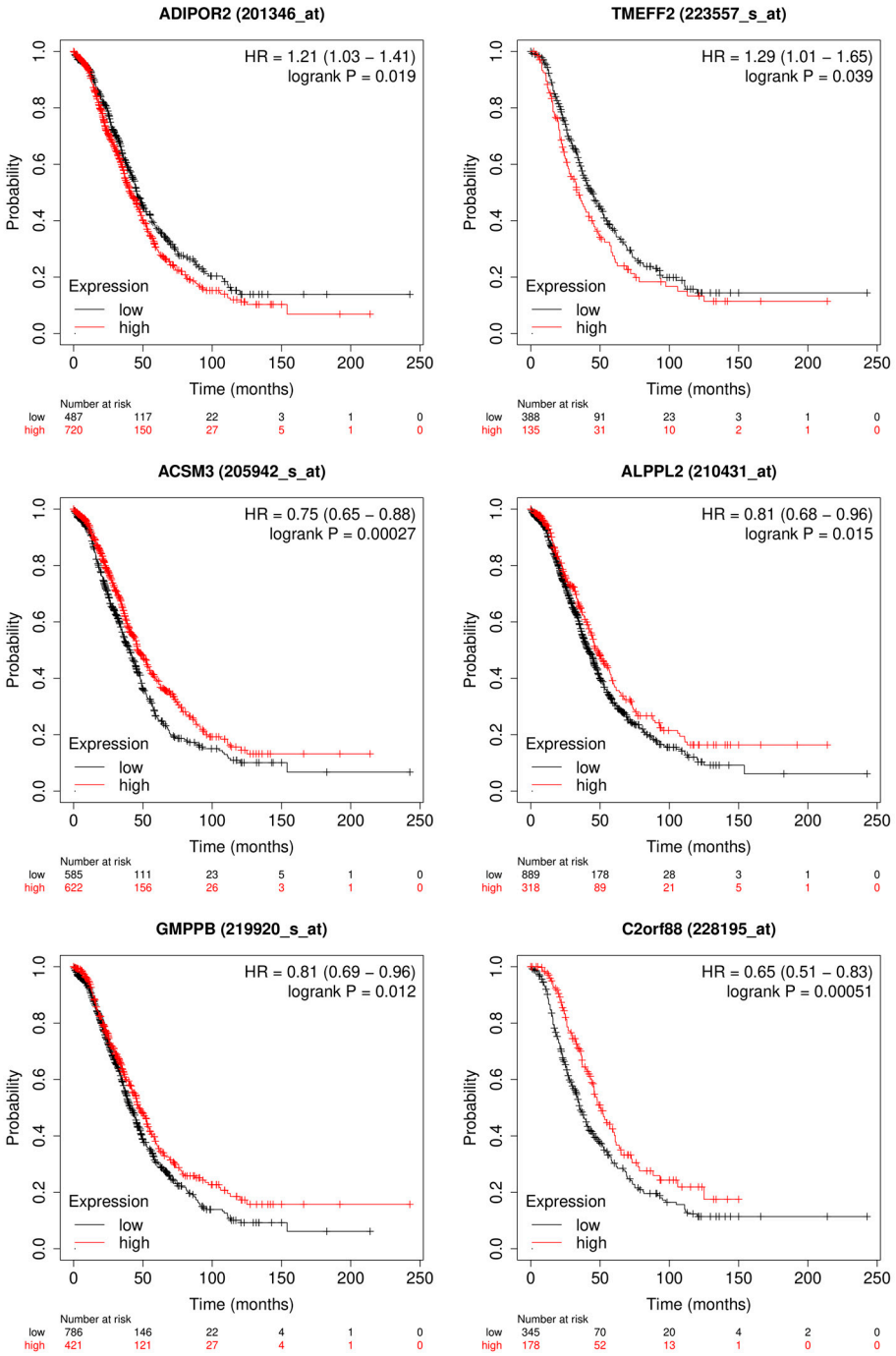


Figure 8. Depicted are the Kaplan-Meier plots for OS of the most relevant genes for the logistic regression model. The genes: ADIPOR2 and TMEFF2 are upregulated in the group of patients with bad outcome. The genes: ACSM3, ALPPL2, GMPPB and C2orf88 are upregulated in the group of patients with good outcome.

3.2. Platinum resistance prediction of ovarian cancer patients

3.2.1. Machine Learning Models Performance

As for the outcome prediction the same six machine learning methods have been used to predict whether a patient is platinum sensitive or resistant. The f1-score is used here as well to assess the performance of the models. The random forest model is the one with the highest f1-score of 0.91 and two misclassifications from 38 observations. The logistic regression model has an f1-score of 0.89 and two misclassifications. The random forest model used the data from the PCA and is therefore more

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2023 doi:10.20944/preprints202305.0413.v1

difficult to interpret. For the feature analysis with SHAP the logistic regression model is used instead. The K-means clustering model is here the worst performing model as well. For the selection of the genes the models have been tested on the differentially expressed genes with a p-adjusted value of 0.1 and 0.05. The gene dataset with a cutoff of 0.05 consisting of 172 genes was the best performing one.

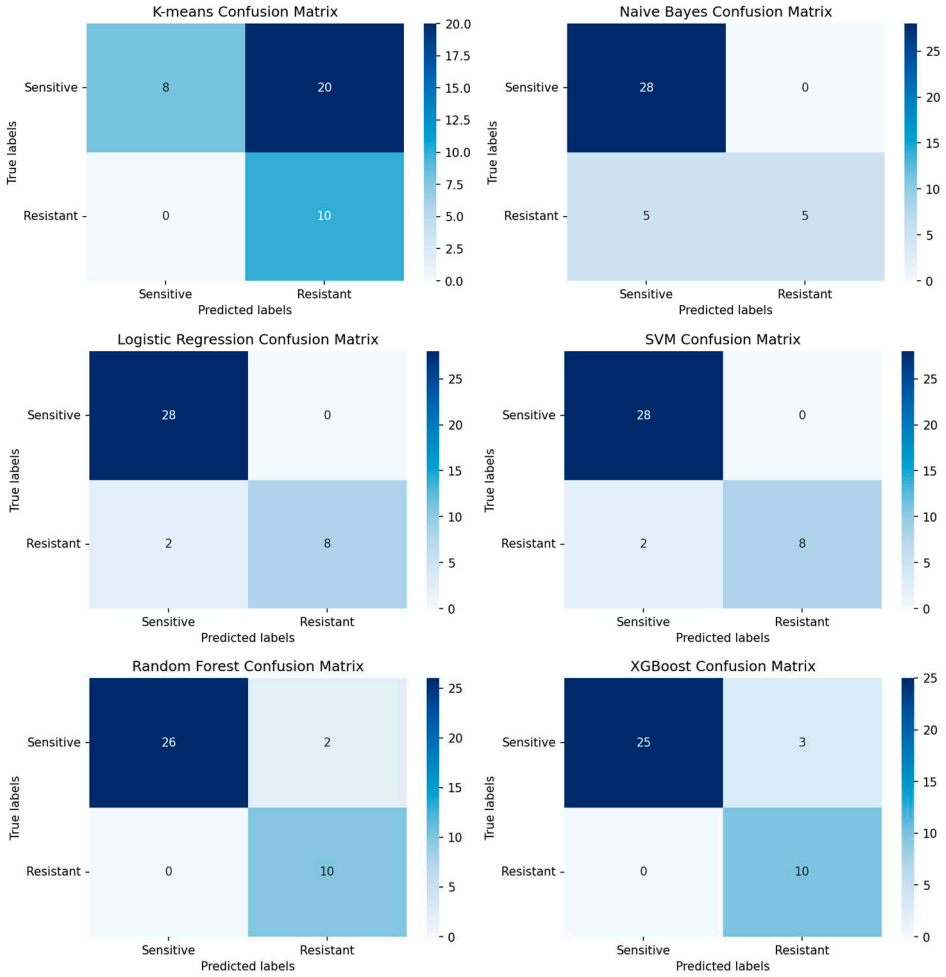


Figure 9. Depicted confusion matrixes for the different machine learning methods which have been used. The higher the number for the corresponds between true labels and predicted labels the more accurate and trustworthy the model is. Following are the f1-scores that evaluates the performance of the model: K-means Clustering: 0.5, Naïve Bayes: 0.67, Logistic Regression: 0.89, Supported Vector Machine: 0.89, Random Forest: 0.91, XGBoost: 0.87. The random forest model is the model with the highest f1-score and two misclassifications from 38 observations. For higher interpretability the logistic regression model is used for further analysis.

Table 2. Overview of the performance of machine learning models for the platinum resistance status prediction of ovarian cancer patients.

Machine learning Method	f1-score
K-means clustering	0.5
Naive Bayes	0.67
SVM	0.89
Logistic Regression	0.89
Random Forest	0.91
XGBoost	0.87

 doi:10.20944/preprints202305.0413.v1

3.2.2. Identified genes

For the selections of genes the SHAP algorithm has been performed on the logistic regression model to evaluate which genes have the highest impact on the prediction whether the patient is platinum sensitive or resistant. The genes with the highest impact on the model are depicted in Figure 9A. The top 20 genes with the highest Shapley values can be found in the Figure 2A. Those genes were then separated into the ones that are upregulated in the patient group that is platinum sensitive and those that are upregulated in the platinum resistant group. In Figure 9B the Kaplan-Meier plots are depicted from six genes with high Shapley values and low adjusted p-values. The first two are upregulated in the group of patients with bad outcome and the other four are upregulated in the group of patients with good outcome.

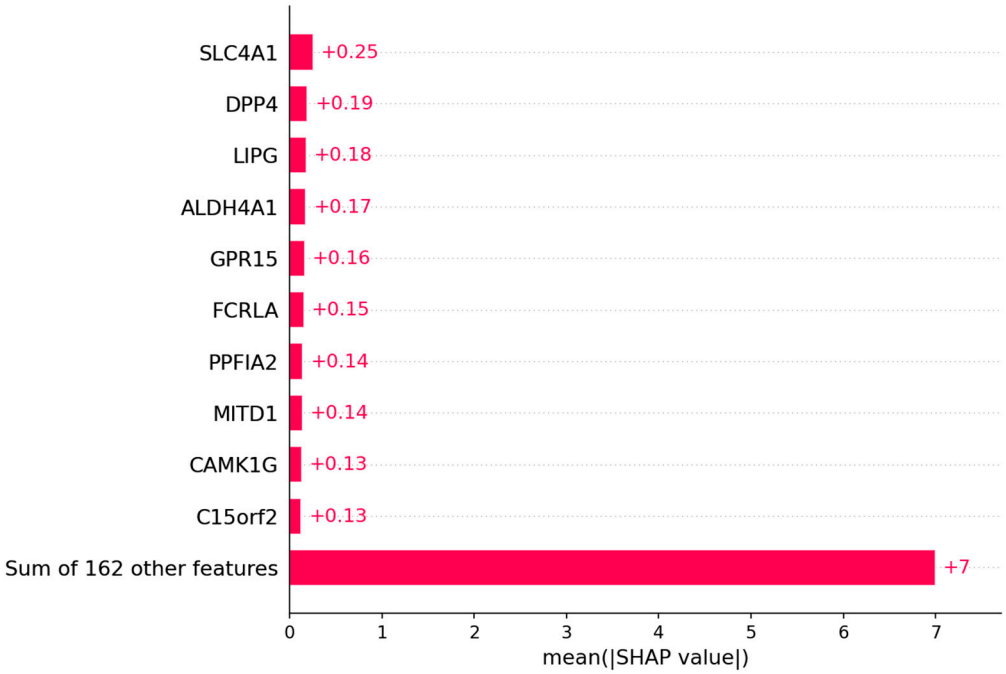


Figure 10. The bar plot shows the top 10 genes based on the mean SHAP value of the logistic regression model of the platinum resistance prediction.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2023 doi:10.20944/preprints202305.0413.v1

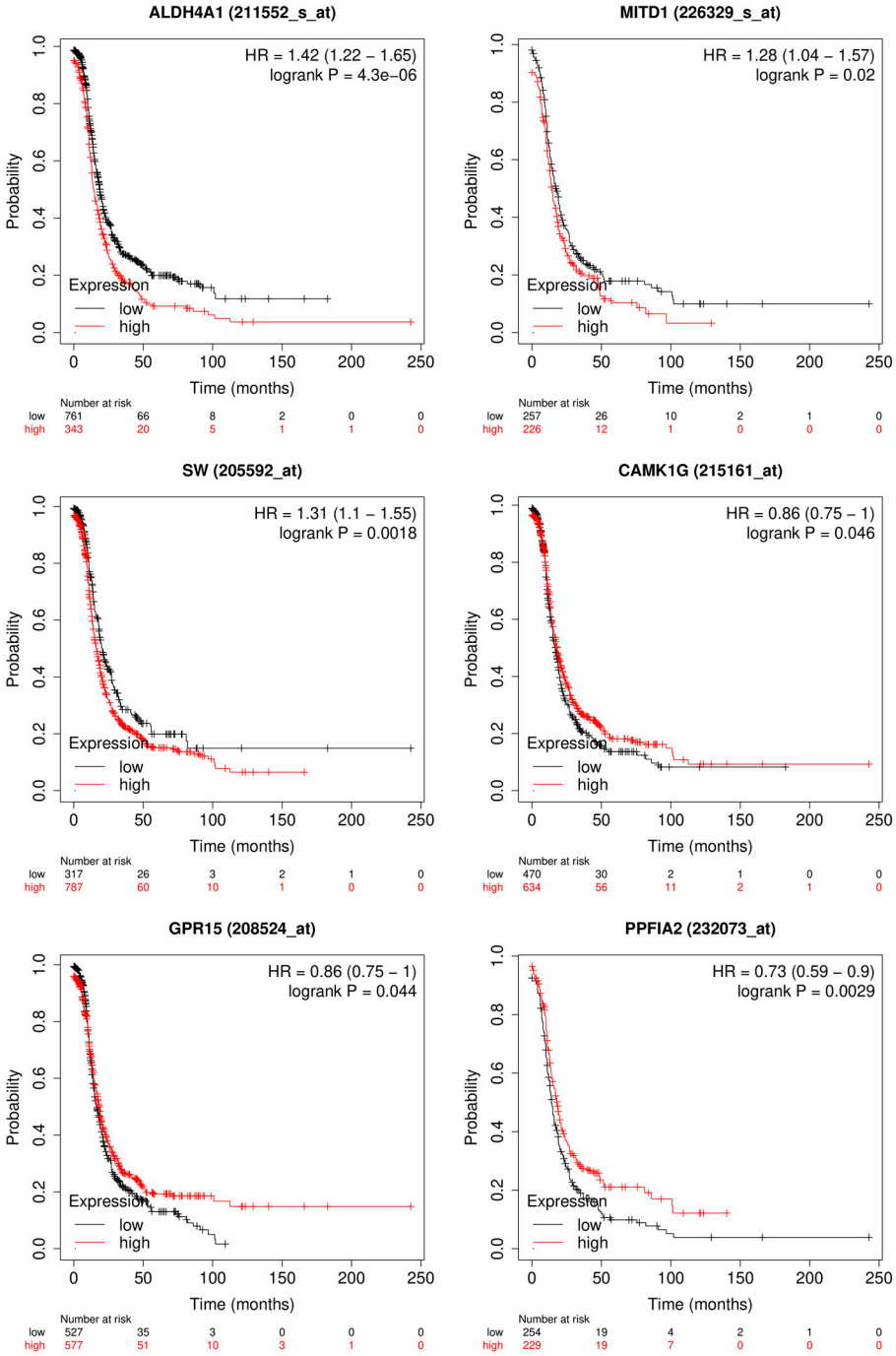


Figure 11. Depicted are the Kaplan-Meier plots for progression free survival (PFS) of the most relevant genes for the logistic regression model. The genes: ALDH4A1, MITD1 and SLC4A1 (SW) are upregulated in the patients with a platinum resistance. The genes: CAMK1G, GPR15 and PPFIA2 are upregulated in the patients that are platinum sensitive.

4. Discussion

4.1. Outcome of the patients

For the outcome prediction of ovarian cancer patients based on their gene expression profile it was possible to demonstrate that you can reach a high performance with common machine learning methods. Apart from the K-means clustering method all machine learning models had a decent prediction performance. As shown in Figure 6 in the confusion matrixes the models were able to

 doi:10.20944/preprints202305.0413.v1

predict good and bad outcome very well. This is probably due to the data preprocessing steps and filtering of the genes based on their differential expression. With 149 genes from originally 20429 the data size has been decreased significantly and also the genes with none or low information value have been removed. Furthermore the separation of the patients between those that have a very short survival with 2 years or less after their treatment and those with long survival of 5 years and longer probably increased the differences in gene expression as well. Looking at the Kaplan-Meier plots in Figure 8 it was possible to identify genes that are significant for the OS of ovarian cancer patients using the SHAP algorithm. TMEFF2 is the gene with highest median Shapley value. It has been shown that high expression of TMEFF2 in endometrial cancer is correlated with advanced cancer stage, poor differentiation and lymph node metastasis [35]. Expression is also correlated with the recurrence of the tumor after successful therapy [36]. ADIPOR2 is in the top 20 of the selected genes by SHAP and is correlated as shown in the Kaplan-Meier plot in Figure 8 with shorter OS survival when highly expressed. It has been shown in chicken ovarian cancer cell lines that ADIPOR2 protein is significantly higher expressed in cancerous ovaries than in normal ovaries [37]. ACSM3 is number four of the median Shapley values. High expression of this gene on the other hand is correlated with inhibited cell proliferation, migration and invasion of ovarian cancer cells. Overexpression of the gene even led to suppression in cell migration [38]. Moreover in High-grade serous ovarian carcinoma (HGSOC) ACSM3 is able to suppress tumor growth in vitro and in vivo [39]. Even though the biological implications of ALPPL2 remains unclear high expression of the gene are correlated with good outcome of patient as shown in Figure 8 and it has been reported as true tumor specific antigen [40]. For GMPPB it is a similar case. The increased expression of it is correlated with favorable outcome. There is little information about its role in cancer but 2 studies identified the gene as a predictive marker in ovarian cancer as well [41,42]. The last gene shown is C2orf88 which is upregulated in patients with longer OS. Regarding the biological function C2orf88 it is only predicted that it enables protein kinase A regulatory subunit binding activity [43]. But it is definitely a prognostic factor for ovarian cancer patients. The 2 most promising genes to function as biomarker and as prognostic factor are TMEFF2 and ACSM3 due to their high significance and known biological functions.

4.2. Prediction of platinum resistance status

It was possible to create multiple well performing machine learning models to predict whether a patient is platinum sensitive or resistant. The highest score reached the random forest model with an f1-score of 0.91. The principal components from the PCA have been used as input. In this paper they used a similar approach and trained a deep learning model with a much bigger patient cohort with 2616 samples. The best performance of their model in predicting the platinum resistance status of patients had an f1-score of 83.1. The higher performance of the random forest model used here could be due to the different data preprocessing or the smaller sample size [44]. In Figure 11 the Kaplan-Meier plots of 3 genes that are upregulated in patients with platinum sensitivity and 3 genes that are upregulated in patients with resistance are depicted. In comparison to the outcome prediction the plots in Figure 11 show the PFS of the ovarian cancer patients instead. The reason for it is that patients that have a recurrence within 6 months after treatment are considered resistant and patients that have no recurrences or one after 6 months are considered sensitive. Therefore the PFS is a better fit to identify genes that are significant for a patient acquiring resistance or being sensitive to platinum. SLC4A1 (SW) has the highest median Shapley value and it's increased expression is correlated with low PFS. The gene is upregulated in patients that are resistant to platinum. As stated in another paper the gene is an independent for poor OS in grade ¾ serous ovarian cancer [45]. The protein AE1 is a chloride/bicarbonate transporter which is encoded by SLC4A1. AEs are important to regulate the intracellular pH [46]. Alterations in pHi are frequently altered in different types of cancer, like ovarian cancer [47,48]. ALDH4A1 has the fourth highest median Shapley value and the high expression of it is highly correlated with platinum resistance and poor PFS. The gene has been associated by other studies with chemoresistance and might mediate carboplatin resistance [49,50]. MITD1 is the last one of the platinum resistant group and the high expression of this gene is

 doi:10.20944/preprints202305.0413.v1

associated with low PFS as well [51]. The protein coded by MITD1 is recruited by ESCRT-III is recruited to midbodies and participates afterwards cytokinesis abscission [52]. ESCRT-III has been shown to be in disorder in ovarian cancer and therefore the higher expression of MITD1 might have a negative effect on that [53]. High expression of CAMK1G is associated with longer PFS survival and it is upregulated in the patients that are platinum sensitive. The gene encodes with 3 other genes the protein kinase I family. These enzymes control a wide range of functions in cancer and might be potential therapy targets [54]. GPR15 has the same prognostic attribution as CAMK1G. It has been shown that GPR15 has the potential with his natural ligand to inhibit cancer cell growth [55]. High gene expression of PPFIA2 is highly correlated with longer PFS in ovarian cancer patients. Not too much research has been done to identify the correlation between PPFIA2 and cancer types but the protein it encodes binds to calcium/calmodulin dependent kinases [56]. It has already been suggested that Ca^{2+} signaling is important in cancer cell function so there might be a correlation between Ca^{2+} pathways and acquiring platinum resistance.

5. Conclusions

It has been demonstrated in this approach that it is possible to predict the outcome and resistance status of ovarian cancer patients and identify biological relevant genes. The most promising potential biomarkers are: TMEFF2, ACSM3, SLC4A1 and ALDH4A1. Their SHAP median values were high, they had a strong correlation with OS or PFS and their biological functions affect the cancer.

Author Contributions: Conceptualization, Vincent Schilling, Peter Beyerlein and Jeremy Chien; Data curation, Vincent Schilling; Formal analysis, Vincent Schilling; Funding acquisition, Jeremy Chien; Investigation, Vincent Schilling and Jeremy Chien; Methodology, Vincent Schilling, Peter Beyerlein and Jeremy Chien; Project administration, Jeremy Chien; Resources, Jeremy Chien; Software, Vincent Schilling and Jeremy Chien; Supervision, Peter Beyerlein and Jeremy Chien; Validation, Vincent Schilling and Jeremy Chien; Visualization, Vincent Schilling; Writing – original draft, Vincent Schilling; Writing – review & editing, Vincent Schilling, Peter Beyerlein and Jeremy Chien.

Funding: This research received no external funding.

Data Availability Statement: For the analysis the publicly available ovarian cancer data by the TCGA has been used. For further questions on how to obtain the data please contact the authors.

Acknowledgments: We acknowledge Dr. Alice Barr for providing gynecological/oncological expertise.

Conflicts of Interest: The authors declare no conflict of interest, financial or otherwise.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2023 doi:10.20944/preprints202305.0413.v1

Appendix A

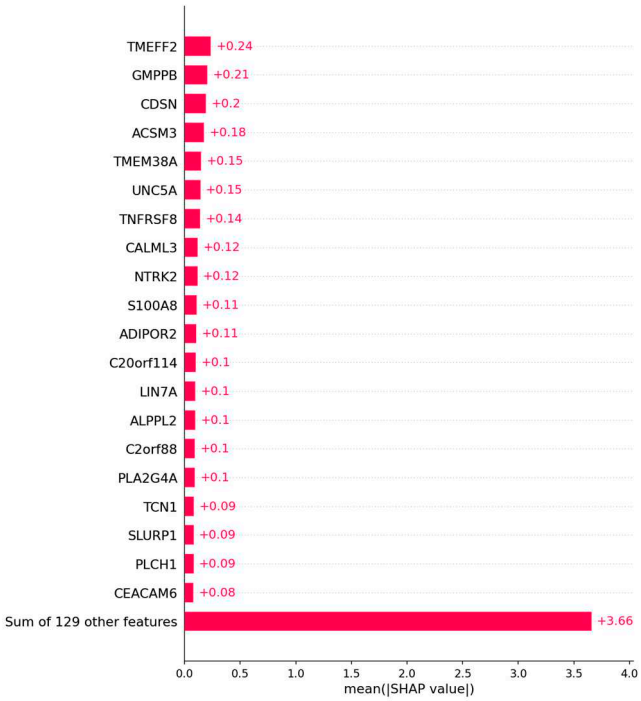


Figure A1. The bar plot shows the top 20 genes based on the mean SHAP value of the logistic regression model of the ovarian cancer outcome prediction.

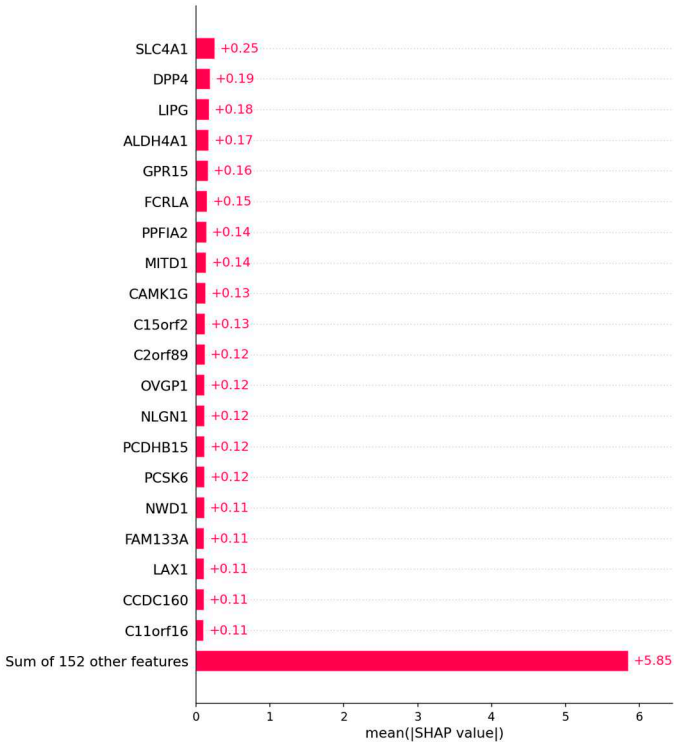


Figure A2. The bar plot shows the top 20 genes based on the mean SHAP value of the logistic regression model of the platinum resistance prediction.

References

1. Ovarian Cancer Survival Rates | Ovarian Cancer Prognosis Available online: https://www.cancer.org/cancer/ovarian-cancer/detection-diagnosis-staging/survival-rates.html (accessed on 26 April 2023).

2. Surgery for Recurrent Ovarian Cancer May Help Selected Patients - NCI Available online: https://www.cancer.gov/news-events/cancer-currents-blog/2022/ovarian-cancer-return-surgery-desktop-iii (accessed on 28 March 2023).
3. Flynn, M.J.; Ledermann, J.A. Ovarian Cancer Recurrence: Is the Definition of Platinum Resistance Modified by PARPi and Other Intervening Treatments? The Evolving Landscape in the Management of Platinum-Resistant Ovarian Cancer. *Cancer Drug Resist* **2022**, *5*, 424–435, doi:10.20517/cdr.2022.13.
4. Jayson, G.C.; Kohn, E.C.; Kitchener, H.C.; Ledermann, J.A. Ovarian Cancer. *The Lancet* **2014**, *384*, 1376–1388, doi:10.1016/S0140-6736(13)62146-7.
5. How to Check for Ovarian Cancer | Ovarian Cancer Screening Available online: https://www.cancer.org/cancer/ovarian-cancer/detection-diagnosis-staging/detection.html (accessed on 28 March 2023).
6. Klein, M.E.; Dabbs, D.J.; Shuai, Y.; Brufsky, A.M.; Jankowitz, R.; Puhalla, S.L.; Bhargava, R. Prediction of the Oncotype DX Recurrence Score: Use of Pathology-Generated Equations Derived by Linear Regression Analysis. *Mod Pathol* **2013**, *26*, 658–664, doi:10.1038/modpathol.2013.36.
7. Kumar, L.; Greiner, R. Gene Expression Based Survival Prediction for Cancer Patients—A Topic Modeling Approach. *PLOS ONE* **2019**, *14*, e0224446, doi:10.1371/journal.pone.0224446.
8. Cardoso, F.; van't Veer, L.J.; Bogaerts, J.; Slaets, L.; Viale, G.; Delaloge, S.; Pierga, J.-Y.; Brain, E.; Causeret, S.; DeLorenzi, M.; et al. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. *New England Journal of Medicine* **2016**, *375*, 717–729, doi:10.1056/NEJMoa1602253.
9. Tang, Z.; Li, C.; Kang, B.; Gao, G.; Li, C.; Zhang, Z. GEPIA: A Web Server for Cancer and Normal Gene Expression Profiling and Interactive Analyses. *Nucleic Acids Res* **2017**, *45*, W98–W102, doi:10.1093/nar/gkx247.
10. Mahood, E.H.; Kruse, L.H.; Moghe, G.D. Machine Learning: A Powerful Tool for Gene Function Prediction in Plants. *Appl Plant Sci* **2020**, *8*, e11376, doi:10.1002/aps3.11376.
11. Johnsen, P.V.; Riemer-Sørensen, S.; DeWan, A.T.; Cahill, M.E.; Langaas, M. A New Method for Exploring Gene–Gene and Gene–Environment Interactions in GWAS with Tree Ensemble Methods and SHAP Values. *BMC Bioinformatics* **2021**, *22*, 230, doi:10.1186/s12859-021-04041-7.
12. Divate, M.; Tyagi, A.; Richard, D.J.; Prasad, P.A.; Gowda, H.; Nagaraj, S.H. Deep Learning-Based Pan-Cancer Classification Model Reveals Tissue-of-Origin Specific Gene Expression Signatures. *Cancers* **2022**, *14*, 1185, doi:10.3390/cancers14051185.
13. Bi, Y.; Xiang, D.; Ge, Z.; Li, F.; Jia, C.; Song, J. An Interpretable Prediction Model for Identifying N7-Methylguanosine Sites Based on XGBoost and SHAP. *Molecular Therapy - Nucleic Acids* **2020**, *22*, 362–372, doi:10.1016/j.omtn.2020.08.022.
14. Bell, D.; Berchuck, A.; Birrer, M.; Chien, J.; Cramer, D.W.; Dao, F.; Dhir, R.; DiSaia, P.; Gabra, H.; Glenn, P.; et al. Integrated Genomic Analyses of Ovarian Carcinoma. *Nature* **2011**, *474*, 609–615, doi:10.1038/nature10166.
15. Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.M.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.M. The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat Genet* **2013**, *45*, 1113–1120, doi:10.1038/ng.2764.
16. Love, M.I.; Huber, W.; Anders, S. Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biology* **2014**, *15*, 550, doi:10.1186/s13059-014-0550-8.
17. Steinhaus, H. *Bulletin de l'acade´mie polonaise des sciences*. October 19 1956, pp. 801–804.
18. Lloyd, S. Least Squares Quantization in PCM. *IEEE Trans. Inform. Theory* **1982**, *28*, 129–137, doi:10.1109/TIT.1982.1056489.
19. Bayes, T.; Price, null LII. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London* **1763**, *53*, 370–418, doi:10.1098/rstl.1763.0053.
20. Garnier, J.-G.; Quetelet, A. *Correspondance mathématique et physique*; M.Hayez, imprimeur, 1838;
21. Verhulst, P.-F. *NOUVEAUX MÉMOIRES DE L'ACADÉMIE ROYALE DES SCIENCES ET BELLES-LETTRES DE BRUXELLES*; L'Académie Royale de Bruxelles et de l'Université Louvain, 1845;
22. Verhulst, P.-F. *MÉMOIRES DE L'ACADÉMIE IMPÉRIALE ET ROYALE DES SCIENCES ET BELLES-LETTRES DE BRUXELLES*; A. BRUXELLES; DE L'IMPRIMERIE ACADÉMIQUE, 1847;
23. Vapnik, V.N.; Lerner, A.Ya. Recognition of Patterns with help of Generalized Portraits. *Recognition of Patterns with help of Generalized Portraits* **1963**, *24*, 774–780.
24. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach Learn* **1995**, *20*, 273–297, doi:10.1007/BF00994018.
25. Ho, T.K. Random Decision Forests. In Proceedings of the Proceedings of 3rd International Conference on Document Analysis and Recognition; August 1995; Vol. 1, pp. 278–282 vol.1.
26. Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32, doi:10.1023/A:1010933404324.
27. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13 2016; pp. 785–794.

28. Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions 2017.
29. Zhu, A.; Ibrahim, J.G.; Love, M.I. Heavy-Tailed Prior Distributions for Sequence Count Data: Removing the Noise and Preserving Large Differences. *Bioinformatics* **2019**, *35*, 2084–2092, doi:10.1093/bioinformatics/bty895.
30. Robinson, M.D.; Oshlack, A. A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data. *Genome Biology* **2010**, *11*, R25, doi:10.1186/gb-2010-11-3-r25.
31. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority over-Sampling Technique. *J. Artif. Int. Res.* **2002**, *16*, 321–357.
32. Pearson, K. LIII. On Lines and Planes of Closest Fit to Systems of Points in Space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **1901**, *2*, 559–572, doi:10.1080/14786440109462720.
33. Hotelling, H. Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology* **1933**, *24*, 417–441, doi:10.1037/h0071325.
34. Kaplan, E.L.; Meier, P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* **1958**, *53*, 457–481, doi:10.1080/01621459.1958.10501452.
35. Gao, L.; Nie, X.; Zheng, M.; Li, X.; Guo, Q.; Liu, J.; Liu, Q.; Hao, Y.; Lin, B. TMEFF2 Is a Novel Prognosis Signature and Target for Endometrial Carcinoma. *Life Sciences* **2020**, *243*, 116910, doi:10.1016/j.lfs.2019.116910.
36. Alabiad, M.A.; Harb, O.A.; Hefzi, N.; Ahmed, R.Z.; Osman, G.; Shalaby, A.M.; Alnemr, A.A.-A.; Saraya, Y.S. Prognostic and Clinicopathological Significance of TMEFF2, SMOC-2, and SOX17 Expression in Endometrial Carcinoma. *Experimental and Molecular Pathology* **2021**, *122*, 104670, doi:10.1016/j.yexmp.2021.104670.
37. Tiwari, A.; Ocon-Grove, O.M.; Hadley, J.A.; Giles, J.R.; Johnson, P.A.; Ramachandran, R. Expression of Adiponectin and Its Receptors Is Altered in Epithelial Ovarian Tumors and Ascites-Derived Ovarian Cancer Cell Lines. *International Journal of Gynecologic Cancer* **2015**, *25*, doi:10.1097/IGC.0000000000000369.
38. Yan, L.; He, Z.; Li, W.; Liu, N.; Gao, S. The Overexpression of Acyl-CoA Medium-Chain Synthetase-3 (ACSM3) Suppresses the Ovarian Cancer Progression via the Inhibition of Integrin B1/AKT Signaling Pathway. *Front Oncol* **2021**, *11*, 644840, doi:10.3389/fonc.2021.644840.
39. Yang, X.; Wu, G.; Zhang, Q.; Chen, X.; Li, J.; Han, Q.; Yang, L.; Wang, C.; Huang, M.; Li, Y.; et al. ACSM3 Suppresses the Pathogenesis of High-Grade Serous Ovarian Carcinoma via Promoting AMPK Activity. *Cell Oncol.* **2022**, *45*, 151–161, doi:10.1007/s13402-021-00658-1.
40. Su, Y.; Zhang, X.; Bidlingmaier, S.; Behrens, C.R.; Lee, N.-K.; Liu, B. ALPPL2 Is a Highly Specific and Targetable Tumor Cell Surface Antigen. *Cancer Res* **2020**, *80*, 4552–4564, doi:10.1158/0008-5472.CAN-20-1418.
41. Liu, J.; Li, S.; Feng, G.; Meng, H.; Nie, S.; Sun, R.; Yang, J.; Cheng, W. Nine Glycolysis-Related Gene Signature Predicting the Survival of Patients with Endometrial Adenocarcinoma. *Cancer Cell International* **2020**, *20*, 183, doi:10.1186/s12935-020-01264-1.
42. Bi, J.; Bi, F.; Pan, X.; Yang, Q. Establishment of a Novel Glycolysis-Related Prognostic Gene Signature for Ovarian Cancer and Its Relationships with Immune Infiltration of the Tumor Microenvironment. *Journal of Translational Medicine* **2021**, *19*, 382, doi:10.1186/s12967-021-03057-0.
43. C2orf88 Chromosome 2 Open Reading Frame 88 [Homo Sapiens (Human)] - Gene - NCBI Available online: https://www.ncbi.nlm.nih.gov/gene/84281#summary (accessed on 26 April 2023).
44. Nasimian, A.; Ahmed, M.; Hedenfalk, I.; Kazi, J.U. A Deep Tabular Data Learning Model Predicting Cisplatin Sensitivity Identifies BCL2L1 Dependency in Cancer. *Computational and Structural Biotechnology Journal* **2023**, *21*, 956–964, doi:10.1016/j.csbj.2023.01.020.
45. Qin, L.; Li, T.; Liu, Y. High SLC4A11 Expression Is an Independent Predictor for Poor Overall Survival in Grade 3/4 Serous Ovarian Cancer. *PLoS One* **2017**, *12*, e0187385, doi:10.1371/journal.pone.0187385.
46. Zhang, L.-J.; Lu, R.; Song, Y.-N.; Zhu, J.-Y.; Xia, W.; Zhang, M.; Shao, Z.-Y.; Huang, Y.; Zhou, Y.; Zhang, H.; et al. Knockdown of Anion Exchanger 2 Suppressed the Growth of Ovarian Cancer Cells via MTOR/P70S6K1 Signaling. *Sci Rep* **2017**, *7*, 6362, doi:10.1038/s41598-017-06472-w.
47. Parks, S.K.; Chiche, J.; Pouysségur, J. Disrupting Proton Dynamics and Energy Metabolism for Cancer Therapy. *Nat Rev Cancer* **2013**, *13*, 611–623, doi:10.1038/nrc3579.
48. Damaghi, M.; Wojtkowiak, J.; Gillies, R. PH Sensing and Regulation in Cancer. *Frontiers in Physiology* **2013**, *4*.
49. Tomita, H.; Tanaka, K.; Tanaka, T.; Hara, A. Aldehyde Dehydrogenase 1A1 in Stem Cells and Cancer. *Oncotarget* **2016**, *7*, 11018, doi:10.18632/oncotarget.6920.
50. Ginestier, C.; Korkaya, H.; Dontu, G.; Birnbaum, D.; Wicha, M.S.; Charafe-Jauffret, E. [The cancer stem cell: the breast cancer driver]. *Med Sci (Paris)* **2007**, *23*, 1133–1139, doi:10.1051/medsci/200723121133.
51. Dong, S.; Hou, D.; Peng, Y.; Chen, X.; Li, H.; Wang, H. Pan-Cancer Analysis of the Prognostic and Immunotherapeutic Value of MITD1. *Cells* **2022**, *11*, 3308, doi:10.3390/cells11203308.
52. Lee, S.; Chang, J.; Renvoisé, B.; Tipirneni, A.; Yang, S.; Blackstone, C. MITD1 Is Recruited to Midbodies by ESCRT-III and Participates in Cytokinesis. *Mol Biol Cell* **2012**, *23*, 4347–4361, doi:10.1091/mbc.E12-04-0292.

53. Nikolova, D.N.; Doganov, N.; Dimitrov, R.; Angelov, K.; Low, S.-K.; Dimova, I.; Toncheva, D.; Nakamura, Y.; Zembutsu, H. Genome-Wide Gene Expression Profiles of Ovarian Carcinoma: Identification of Molecular Targets for the Treatment of Ovarian Carcinoma. *Mol Med Rep* **2009**, *2*, 365–384, doi:10.3892/mmr_00000109.
54. Brzozowski, J.S.; Skelding, K.A. The Multi-Functional Calcium/Calmodulin Stimulated Protein Kinase (CaMK) Family: Emerging Targets for Anti-Cancer Therapeutic Intervention. *Pharmaceuticals (Basel)* **2019**, *12*, 8, doi:10.3390/ph12010008.
55. Wang, Y.; Wang, X.; Xiong, Y.; Li, C.-D.; Xu, Q.; Shen, L.; Chandra Kaushik, A.; Wei, D.-Q. An Integrated Pan-Cancer Analysis and Structure-Based Virtual Screening of GPR15. *Int J Mol Sci* **2019**, *20*, 6226, doi:10.3390/ijms20246226.
56. PPFIA2 PTPRF Interacting Protein Alpha 2 [Homo Sapiens (Human)] - Gene - NCBI Available online: https://www.ncbi.nlm.nih.gov/gene/8499#summary (accessed on 27 April 2023).