Computers in Biology and Medicine 186 (2025) 109681

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Predicting the effectiveness of chemotherapy treatment in lung cancer utilizing artificial intelligence-supported serum *N*-glycome analysis

Rebeka Torok [a], Brigitta Meszaros [a,b], Veronika Gombas [c], Agnes Vathy-Fogarassy [c], Miklos Szabo [d], Eszter Csanky [d], Gabor Jarvas [a], Andras Guttman [a,b,*]

[a] Research Institute of Biomolecular and Chemical Engineering, University of Pannonia, Veszprem, Hungary
[b] Horváth Csaba Memorial Laboratory of Bioseparation Sciences, Research Center for Molecular Medicine, Doctoral School of Molecular Medicine, Faculty of Medicine, University of Debrecen, Debrecen, Hungary
[c] Department of Computer Science and Systems Technology, University of Pannonia, Veszprem, Hungary
[d] Department of Pulmonology, Borsod Academic County Hospital, Miskolc, Hungary

ARTICLE INFO

Keywords:
Lung cancer
N-glycans
Capillary electrophoresis
Chemotherapy
Artificial intelligence-based data analysis

ABSTRACT

An efficient novel approach is introduced to predict the effectiveness of chemotherapy treatment in lung cancer by monitoring the serum N-glycome of patients combined with artificial intelligence-based data analysis. The study involved thirty-three lung cancer patients undergoing chemotherapy treatments. Serum samples were taken before and after the treatment. The N-linked oligosaccharides were enzymatically released, fluorophore-labeled, and analyzed by capillary electrophoresis with laser-induced fluorescence detection. The resulting electropherograms were thoroughly processed and evaluated by artificial intelligence-based classifiers, i.e., utilizing a machine learning algorithm to categorize the data into two (binary) classes. The classifier analysis method revealed a strong association between the structural changes in the N-glycans and the outcomes of the chemotherapy treatments (ROC >0.9). This novel combination of bioanalytical and AI methods provided a precise and rapid tool for predicting the effectiveness of chemotherapy.

1. Introduction

Lung cancer represents a significant global health challenge, ranking as the second most prevalent malignant tumor worldwide [1]. The incidence of the disease is notably high in Hungary, where 10,600 cases were diagnosed in 2022 [2]. On a European scale, the issue is staggering, with 484,000 cases reported in the same year and this number has been increasing since [3]. Considering that lung cancer is the leading cause of malignancy-related deaths due to its high incidence, late-stage discovery, complex etiology, heterogeneity, and aggressive nature, more targeted and personalized treatment approaches are in high demand [4,5]. Certain forms of lung cancer progress quickly, creating an urgent need to start effective treatments. Cancer treatment includes various interventions such as surgery, radiation therapy, immunotherapy, targeted therapy, and several types of chemotherapy [6]. Despite recent advances in oncotherapies, chemotherapy is still one of the most used treatment modalities [7]. These cytotoxic agents are used for both palliative and adjuvant purposes. Biomarkers to predict the effectiveness of chemotherapy would be of great clinical importance since chemotherapy proves ineffective in approximately 70–80 % of the cases [8,9]. Inadequately chosen chemotherapy drugs not only compromise the possibility of recovery but can have serious side effects and represent an economic burden [10]. Oncologists and healthcare providers should carefully plan and monitor the outcome of the chemotherapy treatment to minimize side effects while efficiently targeting the tumor. These treatment plans are mainly dependent on cancer subtypes and possibly existing genetic mutations. Therefore, it is particularly important to be able to assess the effectiveness of chemotherapy after the first session. Conventionally, clinicians use CT scans and MRI to test the efficiency of chemotherapy and to make classifications such as partial-complete response (regression), stable (stationary), or disease progression (progression) [11–13]. However, there is an unmet need for high-precision models to predict the effectiveness of chemotherapy (EoC) of lung cancer patients using specific drugs or drug combinations [14]. Principally, there are three different ways to predict EoC by utilizing 1) anamnesis containing all traditional clinically relevant tests, 2) mono- and multi-omics data, and 3) based on imaging of cells, tissues, or organs

* Corresponding author. Horváth Csaba Memorial Laboratory of Bioseparation Sciences, Research Center for Molecular Medicine, Doctoral School of Molecular Medicine, Faculty of Medicine, University of Debrecen, Debrecen, Hungary.
E-mail address: guttmanandras@med.unideb.hu (A. Guttman).

https://doi.org/10.1016/j.compbiomed.2025.109681
Received 12 August 2024; Received in revised form 5 January 2025; Accepted 12 January 2025

Abbreviations:

AI	artificial intelligence
AUC	Area Under Curve
CBP	carboplatin
CDDP	cisplatin
CGE-LIF	capillary gel electrophoresis with laser-induced fluorescence detection
CI	confidence interval
CRT	chemoradiotherapy
ETO	etoposide
GEM	gemcitabine
PEM	pemetrexed
QDA	Quadratic Discriminant Analysis
ROC	Receiver Operating Characteristic
SFS	Sequential Forward Selection
SVC	Support Vector Classifier
SNEC	small cell neuroendocrine carcinoma
TAX	paclitaxel
THF	Tetrahydrofuran
TXT	docetaxel
XGBoost	eXtreme Gradient Boosting

[15–17]. In all these three cases, artificial intelligence (AI) can be intensively utilized to overcome the extremely high complexity of this challenging problem. Recent advancements in using artificial intelligence in EoC prediction are reviewed by Rafique et al. [18]. It is important to note that AI can serve as a classifier specifically tailored to glycan analysis data, but alternative classification algorithms could also be employed. Machine learning (ML) and ML-based classification are important subsets of AI-based tools [19]. The review paper from Patel et al. summarizes numerous challenges in omics data assisted precision medicine and prediction of EoC [20]. Among various biomarkers under investigation, N-glycome analysis has emerged as a promising tool due to its ability to capture subtle yet significant alterations in glycosylation patterns associated with cancer progression and treatment responses [21–23]. The glycan moiety of proteins undergo structural changes in cancerous states that reflect tumor biology and systemic responses [24]. These alterations provide insights into tumor aggressiveness, metastatic potential, and immune evasion mechanisms, establishing their relevance in cancer prognosis [25]. The dynamic nature of glycosylation makes it an ideal candidate for monitoring chemotherapy responses, as it can reflect changes induced by the interventions in real-time. Unlike many biomarkers based on genetic or proteomic-level alterations, N-glycome analysis captures glycosylation changes that mirror dynamic disease states [26,27]. This orthogonal layer of information complements existing tools, enhancing prediction robustness and specificity [28]. Furthermore, AI-based analysis of N-glycosylation profiles allows the detection of subtle, non-linear patterns, often undetectable by conventional statistical methods. For example, while circulating tumor DNA or PD-L1 (programmed death-ligand 1) expression-based models provide valuable insights, they may lack sensitivity or broad applicability.

The main goal of this study is to explore correlations between serum N-glycome changes and the effectiveness of chemotherapy during lung cancer treatment. By employing capillary gel electrophoresis with laser-induced fluorescence detection (CGE-LIF) and ML-based data processing, the relationship between clinical parameters and the alterations of asparagine-linked carbohydrates in serum glycoproteins is investigated. The unique combination of high-resolution N-glycan analysis and state-of-the-art classification tools offer valuable insights into the effectiveness of chemotherapy in lung cancer patients.

2. Materials and methods

2.1. Chemicals and reagents

Sodium dodecyl sulfate (SDS) and Nonidet P-40 were from VWR (Radnor, PA, USA). Acetonitrile, glycerol, dithiothreitol (DTT), tetrahydrofuran (THF), sodium cyanoborohydride (1 M in THF), and acetic acid were from Sigma Aldrich (St. Louis, MO, USA). The Fast Glycan Labeling and Analysis Kit was from Bioscience Kft (Budapest, Hungary). The endoglycosidase PNGaseF for N-glycan release was made in-house as described in Ref. [29].

2.2. Specimen collection

Pathological samples were collected from lung cancer patients undergoing chemotherapy at the Department of Pulmonology, Borsod Academic County Hospital (Miskolc, Hungary), following the appropriate ethical permissions (approval number: 23580-1/2015/EKU (0180/15)) and with informed patient consents. Thirty-three patients of Caucasian descent with lung cancer, receiving diverse doses of chemotherapeutic agents (see detailed information in Table 1), were included in the study. Serum specimens were obtained before and after each treatment session and stored at −80 °C until processing.

2.3. Sample preparation

The sample preparation protocol included denaturation, N-glycan release, fluorophore labeling, and magnetic bead-mediated cleanup. Serum samples were diluted a hundredfold with HPLC-grade water and then denatured at 70 °C for 10 min by adding 2.0 μL of denaturation solution from the Fast Glycan Labeling and Analysis kit (Bioscience Kft). Glycan release was achieved by adding 1.0 μL of PNGaseF enzyme (200 mU) to the reaction mixture followed by incubation at 37 °C for 2 h to ensure complete deglycosylation. The endoglycosidase digestion reaction was stopped by the addition of the labeling solution, which contained 1.0 μL of 40 mM 8-aminopyrene-1,3,6-trisulfonic acid (APTS) in HPLC-grade water, 2.0 μL of $NaBH_3CN$ (1 M in THF), 10 μL of 50 % acetic acid, and 8.0 μL of THF. The reaction mixture was incubated in a heating block overnight at 37 °C in an open vial (all liquid evaporated) [30], purified by using a magnetic bead-based approach, and analyzed by CGE-LIF [31]. All measurements were made in triplicates.

2.4. Capillary gel electrophoresis with laser induced fluorescence detection (CGE-LIF)

A PA800 Plus Pharmaceutical Analysis System with the 32Karat (version 10.1) data collection and processing software package (Beckman Coulter, Brea, CA) was applied for the analysis of the released *N*-linked APTS labeled glycan structures in CGE-LIF mode using 40 cm effective length (50 cm total length), 50 μm ID/365 μm OD bare fused silica capillaries filled with HR-NCHO separation gel buffer (Bioscience Kft). The separations were accomplished by applying 30 kV electric potential in reversed polarity mode (cathode at the injection side, anode at the detection side) at 30 °C capillary temperature. A water plug pre-injection (1.0 psi for 5.0 s) preceded the sample injection by applying 2.0 kV for 2.0 s. Relative percentage area values of the separated peaks were calculated by the Peak Fit v4.12 Software (SeaSolve Software Inc., San Jose, CA). Data quality was ensured by triplicate measurements and rigorous sample handling protocols to minimize the impact of variability and noise in the resulting glycomics data.

2.5. Data analysis

The classifier is a type of machine learning algorithm designed to categorize the data into one or more predefined classes [32]. The classification task, which aimed to explore the correlation between

Table 1
Lung cancer patients undergoing chemotherapy.

Patient	Age	Sex	Histology	Stage	Applied Chemotherapy
1	59	male	squamous cell carcinoma	IV	first-line palliative TAX-CBP
2	55	male	squamous cell carcinoma	IV	first-line palliative GEM-CBP
3	67	male	SNEC	IIIB	first-line palliative CBP-ETO
4	58	female	adenocarcinoma	IV	first-line palliative PEM-CDDP
5	70	male	adenocarcinoma	IV	first-line palliative PEM-CDDP
6	69	male	adenocarcinoma	IV	first-line palliative Bevacizumab + CBP + TAX
7	61	female	adenocarcinoma	IV	first-line palliative GEM-CBP
8	47	male	adenocarcinoma	IIIB	first-line palliative GEM-CBP
9	68	male	adenocarcinoma	IV	first-line palliative GEM-CBP
10	53	male	SNEC	IV	first-line palliative CBP-ETO
11	69	male	adenocarcinoma	IA	adjuvant GEM-CBP
12	58	male	adenosquamosus carcinoma	IIIB	first-line palliative GEM-CBP
13	62	male	SNEC	IV	first-line palliative CDDP-ETO
14	61	male	squamous cell carcinoma	IIIB	first-line palliative GEM-CBP
15	69	female	SNEC	IIIA	first-line palliative CPB-ETO
16	69	female	adenocarcinoma	IIIA	adjuvant GEM + CBP
17	74	male	adenocarcinoma	IIIA	adjuvant GEM + CBP
18	66	female	adenocarcinoma	IV	first-line palliative GEM-CBP
19	63	male	adenocarcinoma	IV	first-line palliative GEM-CBP
20	67	male	squamous cell carcinoma	IIIA	first-line palliative CBP-TXT
21	47	male	adenocarcinoma	IB	adjuvant GEM + CBP
22	62	male	adenocarcinoma	IIIB	first-line palliative GEM-CBP
23	62	male	SNEC carcinoma	IV	first-line palliative CBP-ETO
24	61	male	adenocarcinoma	IB	adjuvant GEM + CBP
25	72	male	squamous cell carcinoma	IV	first-line palliative GEM-CBP
26	56	male	squamous cell carcinoma	IIIB	first-line palliative TAX + CBP
27	65	female	adenocarcinoma	IA	first-line palliative GEM-CBP
28	70	male	squamous cell carcinoma	IV	first-line palliative TAX-CBP
29	63	female	SNEC	IIIA	first-line palliative CBP-ETO
30	73	male	adenocarcinoma	IIIA	first-line palliative GEM-CBP
31	72	female	adenocarcinoma	IV	first-line palliative PEM-CBP
32	59	female	adenocarcinoma	IIIA	CDDP/TXT-CRT
33	65	male	squamous cell carcinoma	IIIB	first-line palliative GEM-CBP

Age average: 63.4, Age median: 63, Age range: 47–74.

chemotherapy outcomes and structural changes in the N-glycome profiles, was transformed into three independent binary classification tasks to predict the effectiveness of chemotherapy, with the following class labels of 'regression,' 'progression,' and 'stationary.' To identify the most suitable classification method, the performance of 27 different classification algorithms was evaluated using their default parameters. In the initial phase, the tested models included linear classifiers (e.g., logistic regression), tree-based classifiers (e.g., decision tree), distance-based classifiers (e.g., k-nearest neighbors), probabilistic Bayes-based models (e.g., Gaussian Naive Bayes, Quadratic Discriminant Analysis (QDA)), support vector classifiers (SVC) with various kernels, ensemble models (e.g., Random Forest, Extra Trees, Gradient Boosting, eXtreme Gradient Boosting (XGBoost)), and a Neural Network-based classifier [33]. Following the performance evaluation with the default parameters, we conducted an exhaustive hyperparameter tuning for the five best-performing machine learning algorithms: SVC, QDA, Random Forest, XGBoost, and Neural Network. For each algorithm, the hyperparameters that most significantly influenced learning were meticulously tuned, as follows. For SVC, hyperparameter tuning focused on the type of kernel applied. In the case of QDA, the regularization parameter was optimized. For the Random Forest algorithm, the hyperparameter set included the splitting criterion, the number of estimators, the maximum tree depth, the minimum sample size required for a split, and the minimum sample size required at the leaf level. For XGBoost, the tuned hyperparameters were the number of estimators, the maximum tree depth, and the gamma parameter. For the Neural Network, the structure of the network (number of hidden layers, number of neurons) and the activation function were optimized. Hyperparameter tuning was conducted using Bayesian Optimization, with a 5-fold cross-validation applied in all cases. Quadratic Discriminant Analysis showed the best classification performance, thus it was selected to construct the final model [34]. Additionally, a combination of the Sequential Feature Selection (SFS) procedure [35] and the brute force method was employed to minimize the influence of irrelevant features and identify the relevant N-glycan peaks with structural changes most effectively correlating the chemotherapy response. Due to the limited number of records in the original dataset, the fine-tuned QDA classifier was run 1000 times for the final evaluation. This involved randomly partitioning the entire original dataset into separate training and test sets with an 80 %–20 % ratio, and the resulting performance metrics were calculated by averaging the results of the test sets. All in-house developed data analysis code was implemented in Python using Jupyter Notebook v7.1.1 [36]. We also utilized the Receiver Operating Characteristic (ROC) Area Under Curve (AUC) analysis method, which is a frequently used technique for analyzing the accuracy of diagnostic tests. The ROC curve is the plot of the series of true positive points (sensitivity) against the false positive points (1-specificity). An ideal ROC curve jumps towards the upper left corner of the graph indicating good (AUC >0.8) or great (AUC >0.9) discrimination properties. In other words, the higher AUC value of the ROC curve suggests greater discriminative power.

3. Results and discussion

In this study, the N-glycome of 98 serum samples (33 lung cancer patients, multiple samples collected from each patient depending upon their personal therapeutic needs) were analyzed using the CGE-LIF method to explore any structural changes in their carbohydrate profile during chemotherapy treatments. Serum samples were collected after each treatment session, and the asparagine-linked oligosaccharides were enzymatically released, and labeled with a fluorophore (APTS) for downstream analysis. In our study, we aimed to mitigate the inherent variability in the serum N-glycome profile between individuals by using paired samples from the same individuals before and after chemotherapy. A representative electropherogram from a control healthy human serum sample is shown in Fig. 1. The structural identification utilized direct mining of the GU database entries available in the GUcal v1.1c application linked with the GlycoStore data collection [37]. The selection and numbering of the glycan structures, i.e., peaks in the electropherogram, are based on earlier reports [38]. Shortly, only glycans with greater than 1 % relative peak area are selected for downstream data analysis, meaning 21 peaks in our case.

The relative percentage area of the separated and relevant (>1 %) peaks is calculated from the electropherograms of the serum samples collected before and after the chemotherapy treatments. Subsequently, a

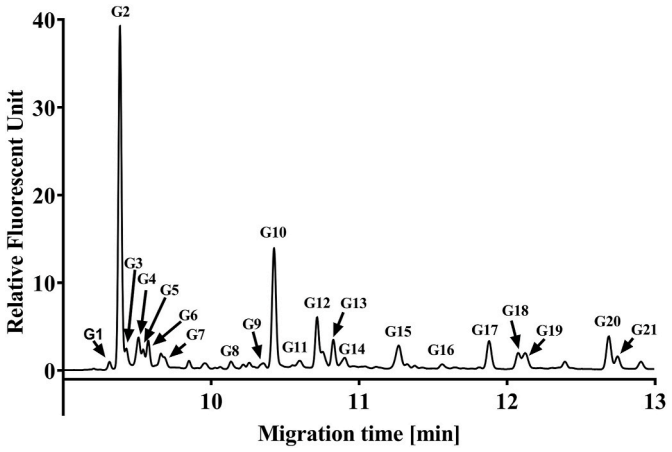

Fig. 1. CGE-LIF separation of PNGase F released and APTS-labeled *N*-glycans from a healthy control human serum sample. The numbered glycans are specified by the Oxford notation [39] for traceability. Peaks: G1: FA4BG4[3,3,3,3]S4, G2: A2G2[6]S2, G3: FA3G3[6]S3, G4: A2G2[3]S2, G5: A2BG2S2, M3, G6: FA2G2S2, G7: FA2BG2S2, G8: FA2[6]G1S1, G9: A3G3[3]S2, G10: A2G2[6]S1, G11: A2BG2S1, G12: FA2G2S1, G13: FA2BG2S1, M7, G14: A4G4[6]S2, G15: FA2, M6, G16: FA2B, G17: FA2[6]G1, M7, G18: FA2[3]G1, G19: FA2B[6]G1, M8, G20: FA2G2, G21: M9. Separation conditions: 40 cm effective capillary length (50 cm total length), 50 μm ID/365 μm OD bare-fused silica capillary; Applied separation voltage: 30 kV (0.17 min ramp time) in reversed polarity mode. LIF detection (excitation: 488 nm/emission: 520 nm); Separation temperature 30 °C. Injection: water pre-injection 5.0 s at 1.0 psi, followed by 2.0 kV/2.0 s sample injection.

QDA classifier is employed to analyze the correlation between changes in the relative peak areas, i.e., changes in the N-glycan profile and the effectiveness of chemotherapy treatment categorized as regression, progression, or stationary. Before classification, the unprocessed input data set was analyzed to explore its consistency. The heterogeneous distribution of the datasets, i.e., True (blue) against False (red) of the three different binary classification tasks is depicted in Fig. 2. Please note, that the term "binary" relates to the type of the classifier, i.e., to distinguish between positive and negative cases only, while the three different tasks are the prediction of regression, progression, and stationary, respectively.

The descriptive power of the N-glycan profile changes is also evaluated using the ANOVA test to shed light on the role of individual peak intensities in the classification tasks. Significant differences are observed for peak IDs G6 and G15 only. In the case of G6, the disparity between the measurements associated with progression and stationary class labels resulted in a p-value of 0.0221. On the other hand, in the instance of G15, the p-value between measurements associated with regression and progression is $p = 0.0091$. Nevertheless, the average probability value calculated using the ANOVA test is notably high with the mean p-value

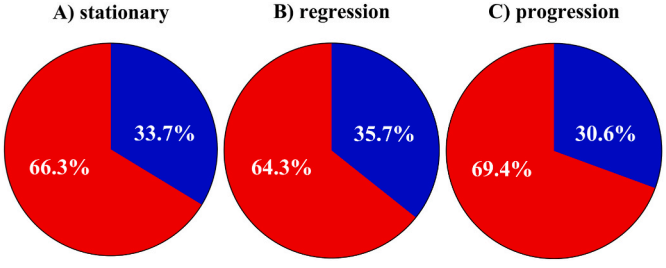

Fig. 2. Pie chart demonstration of the dataset (i.e., class labels) distribution in the case of the three different classification tasks. Approximately 33 % represent True labels (blue), while around 66 % represent False labels (red) for all cases. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

of 0.5003, with a standard deviation of 0.2652. The relative peak intensity distributions of the N-glycan structures grouped by the classification tasks are visualized in Fig. 3.

A thorough inspection of Fig. 3 suggested no significant difference in the distribution of the relative peak intensities in either group, i.e., in regression, progression, or stationary, further confirmed by the ANOVA test above. Furthermore, correlation analysis of the relative percentage area of glycans is performed as well. High correlations are found between the following N-glycan structure pairs: peaks G15/G17 ($r = 0.77$), peaks G17/G18 ($r = 0.91$), peaks G17/G20 ($r = 0.87$), and peaks G18/G20 ($r = 0.77$).

The SFS feature selection method is performed to identify the most relevant glycan structure changes related to chemotherapy response in the binary classification tasks. The SFS algorithm is executed separately for each classification task, i.e., for the prediction of progression, stationary or regression type of chemotherapy response. The results of the SFS method are manually refined using the brute force method. This involved adding and removing glycan structures from the selected set to identify the crucial ones for the classification tasks. From each dataset, the hybrid SFS and brute force algorithms successfully removed the insignificant glycan structures not contributing significantly to the chemotherapy response, resulting in a reduced dataset as listed in Table 2.

As a working hypothesis, we anticipate that all three chemotherapy response types can be accurately predicted using the same data set. Thus, the performance of the QDA classification is first evaluated by using only those glycan structures, which are common in all three cases: peaks G6, G12, G13, G20, and G21. In this instance, the performance of the fine-tuned classification model is not acceptably appropriate as it resulted in the average Area Under Curve (AUC) of 0.6937. This inferior performance of the classifier suggested that the EoC prediction should be handled separately for each response type utilizing the reduced datasets provided by the hybrid SFS and brute force algorithm. Thus, the QDA classifier is employed to evaluate the reduced datasets for each binary classification task. The discriminating power of the model is assessed in terms of the AUC values of the receiver operating characteristic (ROC) curves provided by the QDA classifier.

The fine-tuned QDA classifier is executed on 1000 randomly generated, reduced datasets derived from the original set, to avoid overtraining or convergence to local minima, thus, to provide reliable results. In general, the classifier is trained first, then tested on independent data sets generated by randomly splitting the original data set (80 % training, 20 % test) for each classification task. The AUC values of the averaged 1000 runs resulted in 0.8290 (95 % CI: 0.8200–0.8321) for regression, 0.8295 (95 % CI: 0.8197–0.8323) for progression, and 0.8410 (95 % CI: 0.8356–0.8457) for stationary. Specificity is one of the most challenging criteria in molecular diagnostics and prediction of various diseases. Thus, the specificities are also investigated on the average of a thousand runs resulting in 0.8838 (95 % CI: 0.8793–0.8904), 0.8799 (95 % CI: 0.8711–0.8816), and 0.8535 (95 % CI: 0.8438–0.8555) for regression, progression, and stationary, respectively. The accuracy of the model is 0.7680 (95 % CI: 0.7629–0.7737), 0.7795 (95 % CI: 0.7709–0.7808), and 0.7491 (95 % CI: 0.7447–0.7551) for regression, progression, and stationary, respectively.

Next to the evaluation of the average resulting discriminative power of 1000 randomly generated data sets, the one having the highest AUC values is selected to further demonstrate the clinical potential of the suggested approach. In this instance, the trained classifier was able to achieve AUC values of 0.9981, 0.9231, and 9231 for progression, regression, and stationary, respectively, using the very same data set. The specific ROC curves for each classification task are plotted in Fig. 4. As one can observe, the corresponding specificity values are also rather high, 1 (progression), 0.9231 (regression), and 0.8462 (stationary) suggesting the practical usability of this novel approach. Please note, the surprisingly high prediction power obtained, albeit the utilized data set is restricted by relatively low patient cohort and inhomogeneous

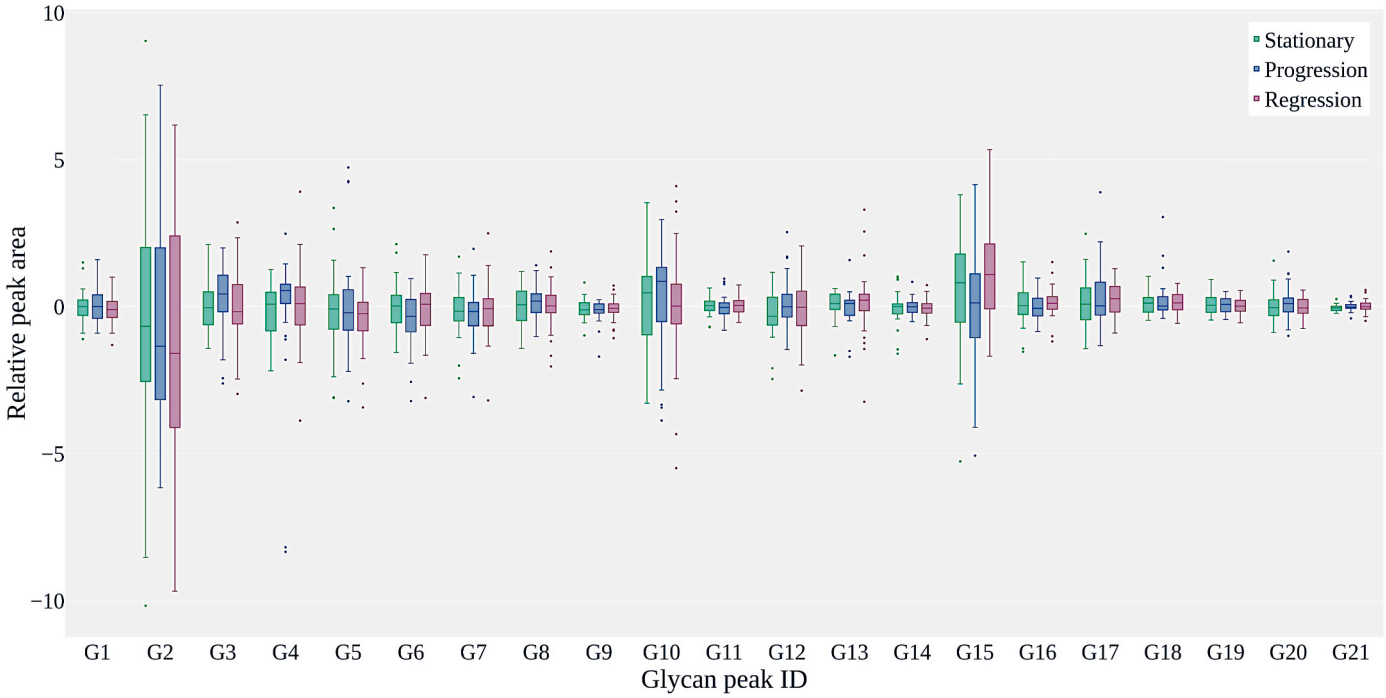

Fig. 3. The relative peak intensity distributions of the N-glycan structures grouped by the classification tasks.

Table 2
The list of the relevant glycan structures for the binary classification tasks.

Classification task	Relevant glycan structure peaks
Regression	G1, G2, G6, G12, G13, G14, G15, G17, G19, G20, G21
Progression	G3, G4, G6, G8, G12, G13, G16, G20, G21
Stationary	G2, G3, G6, G11, G12, G13, G16, G18, G19, G20, G21

consistency with respect to the tumor and therapy types.

4. Conclusion

This study aimed to reveal the correlation between structural changes in the serum N-glycome and chemotherapy treatment response in lung cancer patients, utilizing an artificial intelligence-based

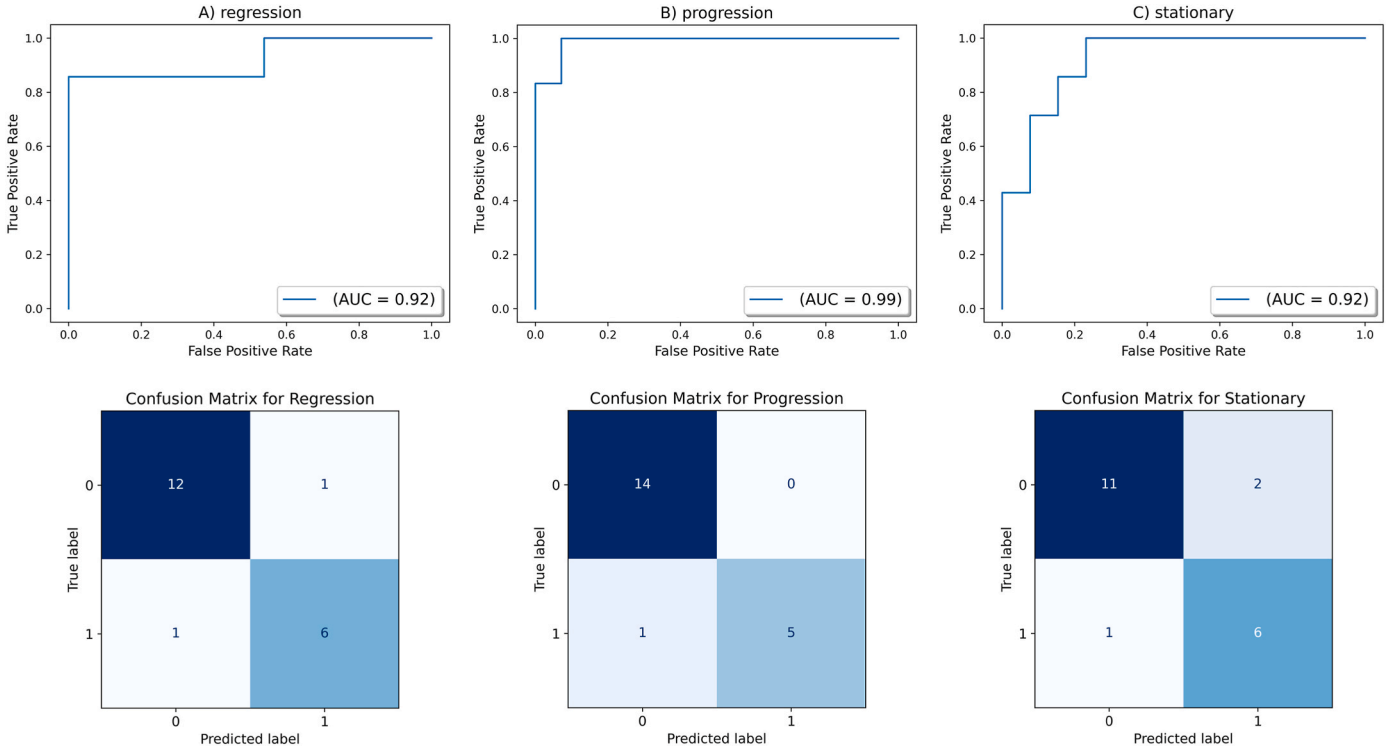

Fig. 4. ROC curves (upper panels) and confusion matrices (lower panels) represent the outstanding discriminative power of the fine-tuned QDA algorithm, in case of prediction of A) regression, B) progression, and C) stationary.

classification model workflow in conjunction with CGE-LIF analysis. A comprehensive analysis of 21 asparagine-linked glycan structures was conducted. Binary classification tasks were established to specifically predict the efficacy of chemotherapy treatment, categorized as regression, progression, or stationary, based on the relative peak area data of the separated glycans. QDA was selected for constructing the final model that underwent fine-tuning to improve the classification accuracy. The combination of the SFS and brute force procedure identified relevant glycan peaks for predicting treatment efficiency across all three binary classification tasks. The resulting AUC values exceeded 0.9, thus, the algorithm successfully predicted the effectiveness of the corresponding chemotherapy, based on changes in the N-glycan profile of serum samples taken before and after the treatment. This synergistic approach not only demonstrated promising results to predict the efficiency of chemotherapy through N-glycan analysis but also highlighted the transformative power of AI in biological sample analysis. A potential limitation of the study is the heterogeneity of the patient population in terms of lung cancer stage, histological type, and the type of chemotherapy treatment used. Additional challenges include the relatively small sample cohort size, the integration of additional clinical data, and improving model interpretability. With our encouraging results, we are planning to expand the study by forming homogeneous groups into a larger patient population.

CRediT authorship contribution statement

Rebeka Torok: Formal analysis. **Brigitta Meszaros:** Visualization. **Veronika Gombas:** Software. **Agnes Vathy-Fogarassy:** Supervision. **Miklos Szabo:** Validation. **Eszter Csanky:** Validation. **Gabor Jarvas:** Supervision, Investigation. **Andras Guttman:** Conceptualization.

Ethics approval and consent to participate

Pathological samples were collected following the appropriate ethical permissions (approval number: 23580-1/2015/EKU (0180/15)) and with informed patient consents.

Consent for publication

Not applicable.

Availability of data and material

The authors confirm that the data supporting the findings of this study are available upon request by email to the corresponding author.

Ethics statement

Pathological samples were collected following the appropriate ethical permissions (approval number: 23580-1/2015/EKU (0180/15); approval date: April 29, 2015) and with informed patient consents.

Funding

Authors gratefully acknowledge the support from the following sources: ATBG Korea V4 joint project of the National Research, Development and Innovation Office of Hungary #2023-1.2.1-ERA_NET-2023-00015, the Andras Koranyi Foundation, the Cooperative Doctoral Program of the Ministry of Culture and Innovation, and by the University of Debrecen Program for Scientific Publication.

Declaration of competing interest

Authors declare no competing interests.

Acknowledgements

This is contribution #211 of the Horváth Csaba Memorial Laboratory of Bioseparation Sciences.

References

[1] C. Li, et al., Global burden and trends of lung cancer incidence and mortality, Chinese Med J 136 (13) (2023) 1583–1590.
[2] K. Bogos, et al., Lung cancer in Hungary, J. Thorac. Oncol. (2020) 1556, 1380 (Electronic)(in eng).
[3] D. T et al., "- The European Cancer Burden in 2020: Incidence and Mortality Estimates for 40," (in - eng), *D - 9005373,* no. - 1879-0852 (Electronic), pp. - 308-347.
[4] K.D. Miller, et al., Cancer treatment and survivorship statistics, CA A Cancer J. Clin. 72 (5) (2022) 409–436, 2022.
[5] J.H. Bi, et al., Observed and relative survival trends of lung cancer: a systematic review of population-based cancer registration data, Thoracic Cancer 15 (2) (2024) 142–151.
[6] Y. Li, B. Yan, and S. He, "Advances and Challenges in the Treatment of Lung Cancer,", no. 1950-6007 (Electronic)(in eng).
[7] C. E. Knezevic and W. Clarke, "Cancer Chemotherapy: the Case for Therapeutic Drug Monitoring," (In Eng), no. 1536-3694 (Electronic).
[8] R. Rosell, M. Cecere F Fau - Santarpia, N. Santarpia M Fau - Reguart, M. Reguart N Fau - Taron, and M. Taron, "Predicting the Outcome of Chemotherapy for Lung Cancer," (In Eng), no. 1471-4892 (Print).
[9] H. Y. Min and H. Y. Lee, "Mechanisms of Resistance to Chemotherapy in Non-small Cell Lung Cancer," (In Eng), no. 1976-3786 (Electronic).
[10] J. H. Schiller et al., "Comparison of Four Chemotherapy Regimens for Advanced Non-small-cell Lung Cancer," (In Eng), no. 1533-4406 (Electronic).
[11] T. Sartoretti, J. E. Wildberger, T. Flohr, and H. Alkadhi, "Photon-counting Detector CT: Early Clinical Experience Review," (In Eng), no. 1748-880X (Electronic).
[12] J.A.-O. Alderuccio, R.A. Kuker, F. Yang, C.H. Moskowitz, Quantitative PET-based biomarkers in lymphoma: getting ready for primetime, Nat. Rev. Clin. Oncol. (2023) 1759–4782 (Electronic).
[13] A. My et al., "- A Lightweight Neural Network with Multiscale Feature Enhancement for Liver CT," (in - eng), *D - 101563288,* no. - 2045-2322 (Electronic), pp. - 14153.
[14] H. Willers, W. L. Azzoli Cg Fau - Santivasi, F. Santivasi Wl Fau - Xia, and F. Xia, "Basic Mechanisms of Therapeutic Resistance to Radiation and Chemotherapy in Lung Cancer," (In Eng), no. 1540-336X (Electronic).
[15] D. M. Bach, W. Straseski Ja Fau - Clarke, and W. Clarke, "Therapeutic Drug Monitoring in Cancer Chemotherapy," (In Eng), no. 1757-6199 (Electronic).
[16] Y. Alduais, H. Zhang, F. Fan, J. Chen, B. Chen, Non-small cell lung cancer (NSCLC): a review of risk factors, diagnosis, and treatment, Medicine 102 (8) (2023) e32899.
[17] H. Hoy, T. Lynch, and M. Beck, "Surgical Treatment of Lung Cancer," (In Eng), no. 1558-3481 (Electronic).
[18] S.M.R.I. Raihan Rafique, Julhash U. Kazi, Machine Learning in the Prediction of Cancer Therapy, – 19, 2021, p. 4017.
[19] B. Mészáros et al., "Machine Learning Based Analysis of Human Serum N-Glycome Alterations to Follow up Lung Tumor Surgery. LID - 10.3390/cancers12123700 [doi] LID - 3700," (in eng), no. 2072-6694 (Print).
[20] B.G. Sk Patel, V. Rai, Artificial Intelligence to Decode Cancer Mechanism: beyond Patient Stratification for Precision Oncology, – 11, 2020.
[21] P. Ss and R. Ca, "- Glycosylation in Cancer: Mechanisms and Clinical Implications," (In - Eng), D - 101124168, no. - 1474-1768 (Electronic), pp. - 540-555.
[22] A. Komaromy, B. Reider, G. Jarvas, and A. Guttman, "Glycoprotein Biomarkers and Analysis in Chronic Obstructive Pulmonary Disease and Lung Cancer with Special Focus on Serum Immunoglobulin G," (In Eng), no. 1873-3492 (Electronic).
[23] D. Thomas, A. K. Rathinavel, and P. Radhakrishnan, "Altered Glycosylation in Cancer: A Promising Target for Biomarkers and Therapeutics," (In Eng), no. 1879-2561 (Electronic).
[24] M. Hires, E. Jane, M. Mego, M. Chovanec, P. Kasak, and J. Tkac, Glycan analysis as biomarkers for testicular cancer. LID - 10.3390/diagnostics9040156 [doi] LID - 156, (in eng), no. 2075-4418 (Print).
[25] B. Reider, G. Jarvas, J. Krenkova, and A. Guttman, "Separation Based Characterization Methods for the N-Glycosylation Analysis of Prostate-specific Antigen," (In Eng), no. 1873-264X (Electronic).
[26] S. Schmid et al., "How to Read a Next-Generation Sequencing Report-What Oncologists Need to Know," (In Eng), no. 2059-7029 (Electronic).
[27] B. Mészáros et al., "Comparative Analysis of the Human Serum N-Glycome in Lung Cancer, COPD and Their Comorbidity Using Capillary Electrophoresis," (In Eng), no. 1873-376X (Electronic).
[28] K. A.-O. Schjoldager, Y. A.-O. Narimatsu, H. A.-O. Joshi, and H. A.-O. Clausen, "Global View of Human Protein Glycosylation Pathways and Functions," (In Eng), no. 1471-0080 (Electronic).
[29] R. Farsang, Z. Kovacs, G. Jarvas, A. Guttman, Ultrahigh-sensitivity capillary electrophoresis analysis of trace amounts of nitrate and nitrite in environmental water samples, Separations 9 (11) (2022) 333.
[30] B. Reider, M. Szigeti, A. Guttman, Evaporative fluorophore labeling of carbohydrates via reductive amination, Talanta 185 (Aug 1 2018) 365–369, https://doi.org/10.1016/j.talanta.2018.03.101.

[31] V. C, L. C, and G. A, "- Rapid Magnetic Bead Based Sample Preparation for Automated and High Throughput," (in - eng), *D - 0370536,* no. - 1520-6882 (Electronic), pp. - 5682-5687.
[32] S. Ih and O. Id, "- Machine Learning: Algorithms, Real-World Applications and Research Directions," (in - eng), *D - 101772308,* no. - 2661-8907 (Electronic), pp. - 160.
[33] G. Cerulli, Fundamentals of Supervised Machine Learning: with Applications in Python, Springer Nature, 2023.
[34] A. Tharwat, Linear vs. quadratic discriminant analysis classifier: a tutorial, Int. J. Applied Pattern Recognition 3 (2016).
[35] T. Schüppstuhl, K. Tracht, J. Rossmann, Tagungsband des 4. Kongresses Montage Handhabung Industrieroboter, Springer Berlin Heidelberg, 2019.
[36] D. Toomey, Jupyter for Data Science, 2017, p. 242.
[37] J. G, S. M, and G. A, "- GUcal: an Integrated Application for Capillary Electrophoresis Based Glycan," (In - Eng), D - 8204476, no. - 1522-2683 (Electronic), pp. - 3094-3096.
[38] M. B, et al., "- Comparative analysis of the human serum N-glycome in lung cancer, COPD and their,", J. Chromatogr. B 1137 (2020).
[39] H. Dj, M. Ah, R. L, C. Mp, and R. Pm, "- Symbol Nomenclature for Representing Glycan Structures: Extension to Cover," (in - eng), *D - 101092707,* no. - 1615-9861 (Electronic), pp. - 4291-4295.