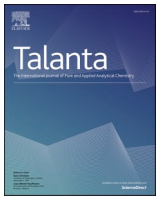
Talanta 295 (2025) 128389

Contents lists available at ScienceDirect

Talanta

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



Improved analytical workflow towards machine learning supported N-glycomics-based biomarker discovery

Agnes Vathy-Fogarassy [a,*], Veronika Gombas [a], Rebeka Torok [b], Gabor Jarvas [b], Andras Guttman [b,c]

[a] Department of Computer Science and Systems Technology, University of Pannonia, Egyetem u 10., Veszprem, H-8200, Hungary
[b] Research Institute of Biomolecular and Chemical Engineering, University of Pannonia, Veszprem, H-8200, Hungary
[c] Horvath Csaba Memorial Laboratory of Bioseparation Sciences, Research Center for Molecular Medicine, Doctoral School of Molecular Medicine, Faculty of Medicine, University of Debrecen, Debrecen, H-4032, Hungary

ARTICLE INFO

Keywords:
Lung cancer
N-glycome
Capillary electrophoresis
Feature selection
Machine learning

ABSTRACT

The composition and function of glycans are very complex thus manual data interpretation of their structural elucidation is difficult. Capillary electrophoresis is one of the liquid phase separation techniques, which is most frequently used to address these challenging tasks. Combining high-resolution capillary electrophoresis with machine learning-supported data interpretation holds the promise to gain as much chemical and clinical information from the analyzed samples as possible. However, this combination requires significant technological improvements both in the analytical and the data processing aspects. In this study we report on the development of an automated, liquid-handling robot-based sample preparation method to obtain reproducible and N-glycome profiles by capillary electrophoresis for the subsequent machine learning-supported data interpretation, which was optimized for the special needs of the analysis. The resulting new glycoanalytical workflow was then tested for a demanding problem to predict the effectiveness of chemotherapy treatments of lung cancer patients ensuring the effective management of the disease. Our findings revealed that the achieved N-glycan data contained important clinical information to accurately predict patient response to chemotherapy with AUC values ranged from 0.8290 to 0.8410.

1. Introduction

Glycans are fundamental biomolecules playing crucial roles in various biological processes, including cell signaling, immune modulation, and disease pathogenesis, just to mention the most important ones [1]. In cancer research, altered N-glycan profiles have been observed across multiple cancer types, and the specific glycoforms identified are often considered as potential diagnostic and therapeutic targets [2–4]. However, their structural complexity and heterogeneity pose significant analytical challenges, necessitating the use of high-resolution separation techniques for accurate characterization [4]. Capillary electrophoresis (CE) has emerged as a powerful tool in glycomics research due to its superior separation efficiency, high sensitivity, and reproducibility. Despite these advantages, the manual interpretation of CE-derived glycan profiles remains a limiting factor, restricting the scalability and clinical applicability of glycomics-based biomarker discovery [5].

To fully realize the potential of liquid phase glyocanalytical methods, robust, high-throughput, and standardized sample preparation workflows are essential to minimize variability and ensure reproducibility. Automated liquid handling systems [6] help to address these challenges by enabling precise, parallel sample processing, significantly reducing human error, and enhancing experimental consistency. It also highlights the advantages of automated sample handling in glycomics, such as high-throughput capabilities and minimized chemical exposure [7].

Building on the high efficiency and precision achieved in glycan analysis through automated workflows, the numerical processing and

Abbreviations: AI, artificial intelligence; AUC, area under the curve; CE, Capillary electrophoresis; ML, machine learning; NN, Neural Network; QDA, Quadratic Discriminant Analysis; RF, Random Forest; RFU, relative fluorescent unit; ROC, receiver operating characteristic; SHAP, SHapley Additive exPlanations; SFS, Sequential Forward Selection; SVM, Support Vector Machine; XGBoost, eXtreme Gradient Boosting.

This article is part of a special issue entitled: Frontiers in electrophoresis published in Talanta.

* Corresponding author.
E-mail address: vathy.agnes@mik.uni-pannon.hu (A. Vathy-Fogarassy).

https://doi.org/10.1016/j.talanta.2025.128389
Received 2 March 2025; Received in revised form 20 May 2025; Accepted 25 May 2025
Available online 26 May 2025

evaluation of these complex datasets can be further enhanced by the development of artificial intelligence (AI)-based classification models, opening new possibilities for data interpretation and biomarker discovery. AI-based classification algorithms have gained significant progress in recent years, particularly in medical diagnostics, image recognition, and the analysis of medical imaging techniques [8]. These algorithms excel at analyzing large datasets and detecting complex patterns. A comprehensive review underscored the benefits of various machine learning (ML) algorithms, including supervised, unsupervised, and reinforcement learning, across diverse fields [9]. In medical imaging, deep learning approaches have significantly improved diagnostic accuracy, such as in SPECT myocardial perfusion imaging for heart disease diagnosis [10]. The integration of ML algorithms with analytical workflows has revolutionized glycomics by facilitating data-driven biomarker discovery and disease classification [11]. Recent studies have demonstrated the potential of ML algorithms in identifying glycan-based disease markers, predicting disease progression, and stratifying patients based on glycan signatures. For instance, Rajpal et al. highlighted the role of explainable AI in biomarker discovery, enabling precise classification of breast cancer subtypes through DNA methylation data analysis [12]. Similarly, Demirhan et al. demonstrated the high diagnostic power of ML-supported mass spectrometry-based glycomics [13]. While these studies underscore the promise of ML in glycomics, their successful application relies on the availability of high-quality, reproducible data, which can only be achieved through optimized sample preparation and analytical methodologies.

In this study, we present the development and optimization of a CE-based, automated, liquid-handling robot-assisted sample preparation method with ML-assisted data interpretation for N-glycomics analysis of biological samples. This workflow was designed to ensure highly reproducible N-glycan profiling, facilitate parallel sample processing, and enable seamless integration into large-scale analytical pipelines. To assess its clinical utility, we applied the developed analytical workflow to a critical biomedical challenge: predicting chemotherapy response for lung cancer patients. To the best of our knowledge, no other research group has specifically examined the relationship between N-glycan structures and chemotherapy efficacy in lung cancer, particularly using an integrated approach that combines automated sample preparation, CE-LIF analysis, and AI-based data interpretation. Our study aims to bridge this gap by identifying glycan structures and structure combinations that carry clinically relevant information regarding responsiveness to chemotherapy. During this endeavor, various classification models were fine-tuned and compared, as well as a novel feature interaction analysis method was developed to pinpoint the most informative N-glycan biomarker panels.

2. Material and methods

2.1. Sample preparation and analysis

Serum samples were collected from 33 Caucasian lung cancer patients undergoing chemotherapy at the Department of Pulmonology, Borsod Academic County Hospital (Miskolc, Hungary). Ethical approval (23580-1/2015/EKU) was obtained, and all patients provided informed consent. Samples were collected before and after each chemotherapy session and stored at −80 °C until further processing. Sample preparation was carried out using an automated liquid-handling robot. Serum samples were diluted 100-fold using HPLC-grade water and denatured at 70 °C for 10 min by adding 2.0 μL of denaturation mixture of 0.6 % SDS, 12.5 mM DTT and 0.06 % NP40. N-glycan release was performed by adding 1.0 μL of PNGaseF enzyme (200 mU), followed by incubation at 37 °C for 2 h to ensure complete deglycosylation. After digestion, labeling was achieved using 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 [14]. The reaction mixture was incubated overnight at 37 °C in an open vial, allowing complete evaporation of the liquids. The labeled glycans were purified using a magnetic bead-based cleanup procedure to remove excess reagents and improve signal clarity [15]. All separations were performed using a PA800 Plus Pharmaceutical Analysis System with the 32Karat (version 10.1) software (Beckman Coulter, Brea, CA). A fused silica capillary with a total length of 50 cm (40 cm effective length) and 50 μm ID/365 μm OD was filled with HR-NCHO gel-buffer system (Bioscience Kft., Budapest, Hungary). The separation was conducted under reversed polarity mode by applying 30 kV electric potential The temperature of the separation capillary was controlled at 30 °C. Sample injection was preceded by a water plug pre-injection (1.0 psi for 5.0 s), followed by sample loading at 2.0 kV for 2.0 s. Relative percentage area values of the separated peaks were quantified using the Peak Fit v4.12 Software (SeaSolve Software Inc., San Jose, CA). Each sample was analyzed in triplicate measurements to ensure high reproducibility and minimize variability. Strict sample handling protocols were applied throughout the workflow to maintain data integrity and consistency in glycomics profiling.

2.2. Dataset

To lay the foundation for our work, a database was created based on the capillary electrophoresis results of serum samples from lung cancer patients. The created dataset included the anonymized patient identification code, the relative peak intensities in the electropherograms, and information about the effectiveness of chemotherapy. We selected 21 glycan structures for the analysis based on their abundance and consistent detectability in human serum, which ensures reliable quantification across samples. To account for variability in glycan concentrations, all peak areas were normalized to the total glycan area and expressed as relative area percentages [25]. The N-glycans were specified based on the Oxford notation [1] as follows: 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. The attribute containing information about the effectiveness of chemotherapy (class label) took on three values: 'regression', 'progression', and 'stationary'. In our study, to compensate for the small sample size, the multiclass classification task was converted into the three binary classification tasks (regression, progression, stationary), each aimed at predicting one class of chemotherapy efficacy. The distribution of the class labels was imbalanced in the available dataset. After converting the problem into binary classification tasks, approximately 33 % represented True labels and around 67 % represented False labels for each classification task. Exploratory data analysis was performed on all data, and no noise or outliers were detected. Therefore, no further data preparation activities were necessary.

2.3. Methodology for determining the relevant N-glycan structures

The fundamental requirement for developing a high-quality classifier model is to provide relevant input data and ensure that no noise (irrelevant and false data) is included as input. Feature selection methods aim to identify the most relevant variables for building both supervised and unsupervised machine learning models; however, this selection process encounters several challenges [16,17]. As the effectiveness of chemotherapy treatment is probably indicated by the different glycan peaks or their combinations (panel), we had to identify which N-glycan molecules have predictive power for the defined classification tasks and which do not contribute significantly to the model development.

Determination of the relevant N-glycan structures was carried out separately for each binary classification task by combining Sequential Forward Selection (SFS) [18] with a manual brute force feature combination detection method. First, the SFS feature selection method was

performed separately for each classification task to determine the feature importance values for each N-glycan structure. The SFS method was executed 50 times for each binary classification task, and the feature importance values were averaged separately for each N-glycan structure within each classification task.

Subsequently, a brute force method was run to identify the most relevant N-glycan structures, also considering their combinations. The main goal of this approach was to identify those N-glycans that contribute to the success of the classification task in combination with others.

The brute-force feature selection and feature combination detection method were executed iteratively as follows. The feature list, sorted in descending order based on the feature significance values, resulting from the SFS method, was split into two disjoint feature sets at the midpoint of the list (rounded down to the nearest integer): a *base set* (*B*) and a *set of candidates* (*C*). An iterative process was then run until no new relevant features or feature combinations were found or there were no longer any N-glycan candidates to test. In this iterative process, we searched for those N-glycans in the candidate set that could improve the accuracy (evaluated based on the resulting AUC value) of the model built on the base set, either individually or in combination. First, the AUC value of the model built on the attributes placed in the base set was calculated. Then, the base set was increased by each subset of 1, 2, and 3 elements from the candidate set, and the AUC values of the models based on the extended feature sets were calculated. The impact of the individual features and feature sets on classification was evaluated by comparing the results (the AUC values of the base and extended models), and the importance of the features was defined based on AUC value differences. Those features whose addition substantially increased the accuracy of the model, together with those feature sets that collectively resulted in a greater-than-expected increase in classification accuracy, were identified as relevant feature combinations. The features that could not increase the classification accuracy were denoted as noise and excluded from further investigation. The next phase of the iterative process was performed by considering the identified relevant features and feature combinations as the base set, and the remaining features (excluding the noise features) formed the candidate set. The iterative process continued until the candidate set was empty. Algorithm 1 in the Appendix provides the pseudocode of the implemented brute-force feature selection and feature combination detection method.

2.4. Methodology for developing the N-glycan-based classification model

One of the key questions of the development was determining the most suitable classifier model for classifying the effectiveness of chemotherapy treatments. To answer this question, we first tested 30 classifier methods with their default settings. After evaluating the results of these tests, we selected the most promising methods, which were as follows: Support Vector Machine (SVM) [19], Quadratic Discriminant Analysis (QDA) [20], Random Forest (RF) [21], eXtreme Gradient Boosting (XGBoost) [22], and Neural Network (NN) [23].

Extensive hyperparameter tuning processes were performed on the selected classification models to develop the most accurate one. First, the search space of the hyperparameters to be tuned was determined (see Table 1), and then the classifiers were tuned for each classification task (regression, progression, stationary) separately using the Bayesian Optimization (BayesSearch) method [24] with 5-fold cross-validation. During the tuning of the neural network model, the Adam optimizer was applied up to 100 epochs, using early stopping regularization with the patience parameter set to 5, and the learning rate was equal to 0.001. The training processes were optimized based on the binary cross-entropy loss function. Considering the limited amount of data, all tuning was executed ten times, randomly splitting the entire dataset each time into disjoint training and validation sets with ratios of 80 % and 20 %.

To evaluate the performance of the classification models, the following quality metrics were applied: accuracy, sensitivity, specificity, F1-score, and AUC value. To counterbalance the small sample size, the quality metrics of the fine-tuned models were calculated as the average results of 1000 runs, based on random train-test splits in portions of 80 % and 20 %.

Table 1
The search space of the hyperparameters for each classifier.

Classifier	Parameters/Hyperparameters
SVM	kernel: RBF, poly, sigmoid, linear
QDA	regularization: 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.07, 0.1
Random Forest	criterion: Entropy, Gini # of trees: 10, 20, 30, 40, 50, 100, 150, 200 max depths: 1, 2, 3, 4, 5, 6, 7 min # of samples to split: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 min # of samples at leaf: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
XGBoost	# of trees: 10, 20, 30, 40, 50, 100, 150, 200 max depths: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 gamma: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Neural Network	hidden layers: 1, 2, 3, 4 # of neurons in hidden layers: min value = 1, max value = 30, step = 1 activation function: relu, sigmoid, tanh, exponential

3. Results and discussion

Ensuring the reproducibility of glycan analysis is crucial for biomarker discovery and clinical applications. Therefore, we first evaluated the reproducibility of the automated sample preparation method using a liquid handling robot. The precision of glycan profiling was assessed by analyzing parallel serum sample replicates under identical conditions. Our results confirmed that automation significantly improved consistency by minimizing pipetting errors and variations in enzymatic reactions, which represent common sources of technical bias in manual workflows. Moreover, the ability to process multiple samples in parallel enhances the scalability of glycomics analysis, making it well-suited for large-cohort clinical studies and medical applications. Fig. 1 shows the electropherogram of APTS-labeled N-glycans from a healthy human serum sample. The migration time is plotted on the x-axis, while the relative fluorescent unit (RFU) is displayed on the y-axis. As reported earlier [25], 21 N-glycan structures are suspected having an important role as lung cancer biomarkers, thus the same pool was considered in this study as well. Those glycans were identified as primary targets and their corresponding peaks are numbered accordingly.

First, the precision and robustness of the automated liquid-handling robot were tested using human serum samples from healthy volunteers. Five parallel samples were analyzed with the developed sample preparation method. Fig. 2 compares the representative electropherograms (A–E) obtained from the N-glycan analyses, demonstrating the identical peak patterns observed for all five samples. These results highlight the high reproducibility of the automated workflow, ensuring consistent glycan profiling from the same biological sample. The parallel processing capability of the liquid handling robot allowed for efficient and simultaneous preparation of multiple samples, minimizing human error and enhancing the throughput of glycan analysis in large-scale studies. The developed automated method demonstrated robustness and reliability, making it suitable for biomarker discovery and clinical applications requiring high precision and repeatability.

To identify potential N-glycan biomarkers associated with chemotherapy response in lung cancer patients, we performed a systematic evaluation of the obtained N-glycan patterns. The integration of machine learning-based feature selection enabled the identification of glycan structures that carry important information for the task. The identified glycans and their combinations (panel) were further validated to assess their predictive capability. The developed feature selection and feature combination detection method (Algorithm 1) was applied iteratively on the dataset for each classification task separately. The resulting N-glycans found to be relevant are presented in Table 2 and further details are disseminated elsewhere [26]. The feature combinations that resulted in a greater-than-expected increase in AUC value in

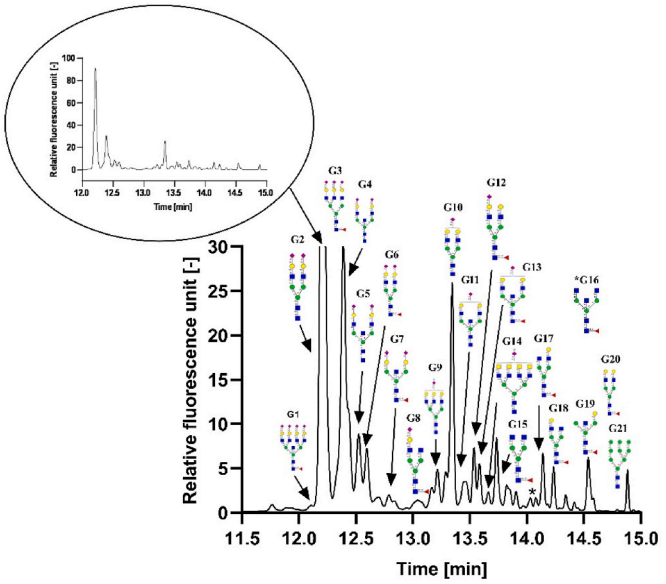


Fig. 1. CE-LIF electropherogram of APTS-labeled N-glycans from a healthy control human serum sample. Peaks: G1: FA4BG4 [3,3,3,3]S4, G2: A2G2 [6]S2, G3: FA3G3 [6]S3, G4: A2G2 [3]S2, G5: A2BG2S2, M3, G6: FA2G2S2, G7: FA2BG2S2, G8: FA2 [6] G1S1, G9: A3G3 [3]S2, G10: A2G2 [6]S1, G11: A2BG2S1, G12: FA2G2S1, 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 parameters: Bare-fused silica capillary with a 50 μm inner diameter and 365 μm outer diameter, featuring a total length of 50 cm and an effective length of 40 cm. The applied separation voltage was set to 30 kV with a ramp time of 0.17 min, operating in reversed polarity mode. LIF detection was performed using an excitation wavelength of 488 nm and an emission wavelength of 520 nm. The separations were conducted at 30 °C. Sample injection included a water pre-injection step (5.0 s at 1.0 psi), followed by sample loading at 2.0 kV for 2.0 s.

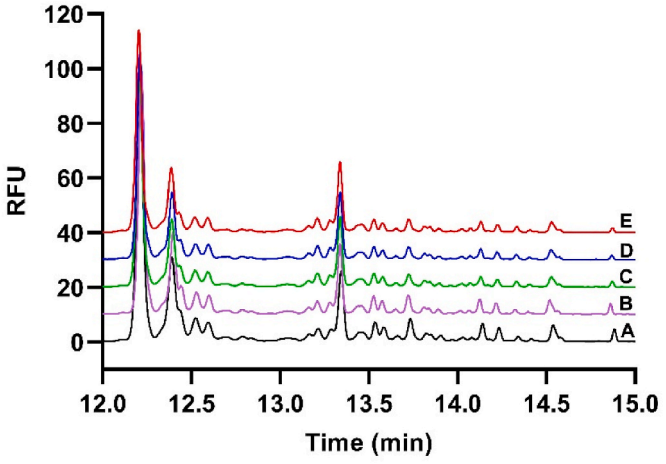


Fig. 2. Representative electropherograms (A–E) obtained from N-glycan analysis using an automated liquid handling robot for sample preparation, demonstrating the high reproducibility and parallel processing capability of the automated workflow from the same biological sample. Conditions were the same as in Fig. 1.

the classification are enclosed in parentheses.

As one can observe, various N-glycan structures were found to be relevant to different classification tasks. The N-glycan structures of G6, G12, G13, G20, and G21 are all present as markers in the disease progression, regression, and stationary state. The G21 glycan structure contributed to the refinement of the model only in combination in all three cases, whereas the G12 was independently discriminative in determining the effectiveness of chemotherapy treatments in all three cases. The N-glycan molecules corresponding to G5, G7, G9, and G10 were deemed irrelevant in assessing the effectiveness of chemotherapy treatment in all three instances.

Table 2
The relevant N-glycan structures for each classification task resulted from the developed feature selection and feature combination detection method.

Classification task	Relevant N-Glycan IDs
Regression	(G1, G2, G6), (G13, G15, G21), G12, G14, G17, G19, G20
Progression	(G3, G8, G13), (G4, G20, G21), G6, G12, G16,
Stationary	(G2, G20), (G3, G13, G16), (G11, G18, G21), G6, G12, G19

To find the best classification method and model, we compared the performance of several classifier models (see Section 2.4). The result of the hyperparameter tuning of the most promising classification methods is summarized in Table 3.

To eliminate possible problems arising from the small sample size, each fine-tuned model was run one thousand times with random training-test splits, and the metrics obtained from the 1000 runs were averaged. This approach was chosen to mitigate the variance and potential biases introduced by any single partitioning of the data. By averaging the results across 1000 runs, we obtained more robust and generalizable performance metrics. Additionally, the computation of standard deviations allowed us to assess the stability of the models, providing insight into how sensitive they are to variations in the training data. The average quality metrics and their standard deviations for the fine-tuned classification models are shown in Table 4. Please note that results are based on unbalanced datasets; therefore, the F1 score and sensitivity are quite small in some cases. Oversampling methods were tested but not applied, as significant improvements in classification accuracy were not observed in the case of its application. In the case of oversampling, although the F1 score and sensitivity were enhanced, the AUC and accuracy values were reduced, most likely due to the small sample size.

It should also be noted that we also compared our results with models that were tuned using feature sets selected by a Sequential Forward Selection algorithm. The results confirmed that the prediction accuracy and AUC values would be better if classifiers were run on the feature sets selected by Algorithm 1.

As shown in Table 4, the QDA classifier achieved high metric values across all classes, indicating its efficacy in predicting the response to chemotherapy treatment for lung cancer. The average AUC and specificity values of the QDA classifier were > 0.82, with average accuracy

Table 3
The hyperparameter values of the fine-tuned classification models.

Classifier	Regression	Progression	Stationary
	Hyperparameters		
SVM	kernel: Gaussian (RBF)	kernel: Gaussian (RBF)	kernel: Gaussian (RBF)
QDA	regularization: 0.001	regularization: 0.001	regularization: 0.001
Random Forest	criterion: Gini; # of trees: 100; max depth: 7; min # of samples to split: 7; min # of samples at leaf: 1	criterion: Gini; # of trees: 20; max depth: 6; min # of samples to split: 4; min # of samples at leaf: 1	criterion: Gini; # of trees: 30; max depth: 6; min # of samples to split: 6; min # of samples at leaf: 1
XGBoost	# of trees: 100; max depth: 6; gamma: 0	# of trees: 50; max depth: 6; gamma: 0	# of trees: 20; max depth: 9; gamma: 0
Neural Network	hidden layer: 2; neurons in hidden layers: 24, 22; act. func.: relu	hidden layer: 3; neurons in hidden layers: 24, 30, 7; act. func.: relu	hidden layer: 1; neurons in hidden layer: 21; act. func.: relu

Table 4
Averaged values of quality metrics of classification models of fine-tuned classifiers on each classification task (based on the attribute sets presented in Table 2).

Classifier	Metric	Regression	Progression	Stationary
SVM	Accuracy:	0.5256 (±0.1026)	0.6927 (±0.0982)	0.5372 (±0.0958)
	AUC:	0.4456 (±0.1248)	0.7544 (±0.1109)	0.4471 (±0.1167)
	F1:	0.3797 (±0.1365)	0.5978 (±0.1150)	0.4454 (±0.1188)
	Sensitivity (recall):	0.4369 (±0.1968)	0.7607 (±0.1695)	0.5473 (±0.1838)
	Specificity:	0.4369 (±0.1624)	0.6636 (±0.1312)	0.5318 (±0.1391)
QDA	Accuracy:	0.7680 (±0.0838)	0.7795 (±0.0818)	0.7491 (±0.0817)
	AUC:	0.8290 (±0.0943)	0.8295 (±0.0993)	0.8410 (±0.0837)
	F1:	0.6120 (±0.1634)	0.5836 (±0.1727)	0.5973 (±0.1492)
	Sensitivity (recall):	0.5531 (±0.1967)	0.5452 (±0.2052)	0.5551 (±0.1855)
	Specificity:	0.8838 (±0.0925)	0.8799 (±0.0910)	0.8535 (±0.0946)
Random Forest	Accuracy:	0.6469 (±0.0653)	0.7064 (±0.0750)	0.5944 (±0.0646)
	AUC:	0.6307 (±0.1335)	0.6910 (±0.1209)	0.5199 (±0.1267)
	F1:	0.2925 (±0.1508)	0.3753 (±0.2089)	0.1839 (±0.1404)
	Sensitivity (recall):	0.2254 (±0.1126)	0.3187 (±0.1864)	0.1427 (±0.1049)
	Specificity:	0.8738 (±0.0940)	0.8726 (±0.0783)	0.8375 (±0.0963)
XGBoost	Accuracy:	0.6224 (±0.0826)	0.7354 (±0.0735)	0.5600 (±0.0406)
	AUC:	0.6104 (±0.1108)	0.7185 (±0.1233)	0.4954 (±0.1172)
	F1:	0.3712 (±0.1675)	0.5177 (±0.1881)	0.2770 (±0.0830)
	Sensitivity (recall):	0.3366 (±0.1423)	0.4983 (±0.1584)	0.2557 (±0.0569)
	Specificity:	0.7762 (±0.1128)	0.8371 (±0.0916)	0.7238 (±0.0663)
Neural Network	Accuracy:	0.6381 (±0.0219)	0.7608 (±0.0743)	0.6393 (±0.0304)
	AUC:	0.6118 (±0.1151)	0.8052 (±0.0873)	0.6343 (±0.1244)
	F1:	0.2520 (±0.1033)	0.5308 (±0.1665)	0.0370 (±0.1206)
	Sensitivity (recall):	0.2007 (±0.0626)	0.4985 (±0.2029)	0.0249 (±0.0746)
	Specificity:	0.8736 (±0.0231)	0.8731 (±0.0817)	0.9702 (±0.0335)

values approaching or exceeding 0.75, suggesting the potential need for further research. The classifier models typically achieved the highest AUC values in predicting disease progression. It is interesting to note that the exception to this finding was the QDA model, which attained high (and nearly identical) AUC values in all three cases. XGBoost, Random Forest, and the Neural Network classifiers performed slightly less effectively, while SVM demonstrated the weakest metrics on this dataset. The fact that the Quadratic Discriminant Analysis method achieved the best result suggests that there is a non-linear relationship between the intensity of the N-glycan peaks in the electropherograms and the effectiveness of the chemotherapy treatment. However, even with a Gaussian kernel, the SVM method could not approach the results achieved by QDA. In the case of applying the neural network model, the poor performance may have resulted from the fact that this approach generally requires large amounts of data for effective learning, which was not available in this study. Overall, we can say that the model created by Quadratic Discriminant Analysis was promising even with such a small amount of data, which certainly suggests a good opportunity for further research.

Additionally, it should be noted that we also developed classification models by excluding the stationary states, since from a clinical perspective, the regression and progression states are more relevant. As a result, the models were developed on an even smaller sample size (65 samples), which represents an additional limitation. Nevertheless, despite the reduced sample size, the results remained consistent. The QDA algorithm performed best, achieving an accuracy of 0.7275 (±0.1150), AUC of 0.8083 (±0.1163), F1 score of 0.7539 (±0.1085), sensitivity of 0.7880 (±0.1564) for the regression class, and specificity of 0.8083 (±0.1952). These results show that, while the sensitivity of the model increased significantly – which is understandable given the cleaned input data – the AUC decreased only slightly despite the substantial reduction in sample size. This further supports our previous observations, suggesting that the selected glycans carry relevant information regarding the effectiveness of chemotherapy treatment.

Fig. 3 shows the inherent potential of applying the resulting QDA model developed for the larger dataset. Here, we present a particularly strong run (though not the best) to illustrate that such results can also occur among multiple runs. This highlights the potential of the method, suggesting that with a larger dataset, even a single run could yield similar results. Although the average quality metrics of the QDA classifier are acceptably high, the trained classifier can achieve much higher AUC values for each classification task using the same dataset (see Fig. 3).

Since the QDA algorithm and the model it generated showed the best results, we used this model in further study. To more deeply investigate the role of individual N-glycans in the classification tasks, a SHAP (SHapley Additive exPlanations) analysis was performed [27]. The SHAP analysis quantifies the contribution of features to the classification prediction. In our case, SHAP values were calculated for the relevant feature sets (see Table 2) to determine the impact of N-glycans.

The SHAP values were calculated as averaged values across 1000 runs with random training-test splits by applying the fine-tuned QDA classification model. The results are shown in Fig. 4.

When evaluating SHAP values, it should be noted that their calculation was based solely on the extent to which a given feature contributed to the accuracy of the classifier model on its own, and the calculation did not consider combinations.

Analyzing the results, we can see that the relevant N-glycans not only varied among different classification scenarios but also differed in their contributions to prediction across various classification cases. Considering the G12 structure, which was identified as an independent biomarker in all three cases, one can see that it contributed significantly more to the accuracy of the model in predicting disease progression and stationary state. In contrast, it had a lower influence in the case of regression.

In Fig. 5 (a), the beeswarm plot illustrates the extent of the importance of individual features (N-glycans) to the accuracy of classifying lung cancer progression in the case of a randomly split dataset. It is shown that, in general, there is no linear relationship between the relative peak intensities of N-glycans and the SHAP values, except for G6. For this feature, it was observed across multiple different splits that higher values (marked in red) contributed less to the prediction, while lower values (marked in blue) enhanced the classification accuracy.

The dependence plot in Fig. 5(b) reveals this relationship more clearly. The SHAP values for the N-glycan G6 are depicted on the y-axis, and the x-axis represents their values. The connection between these two values is clearly visible in the plot.

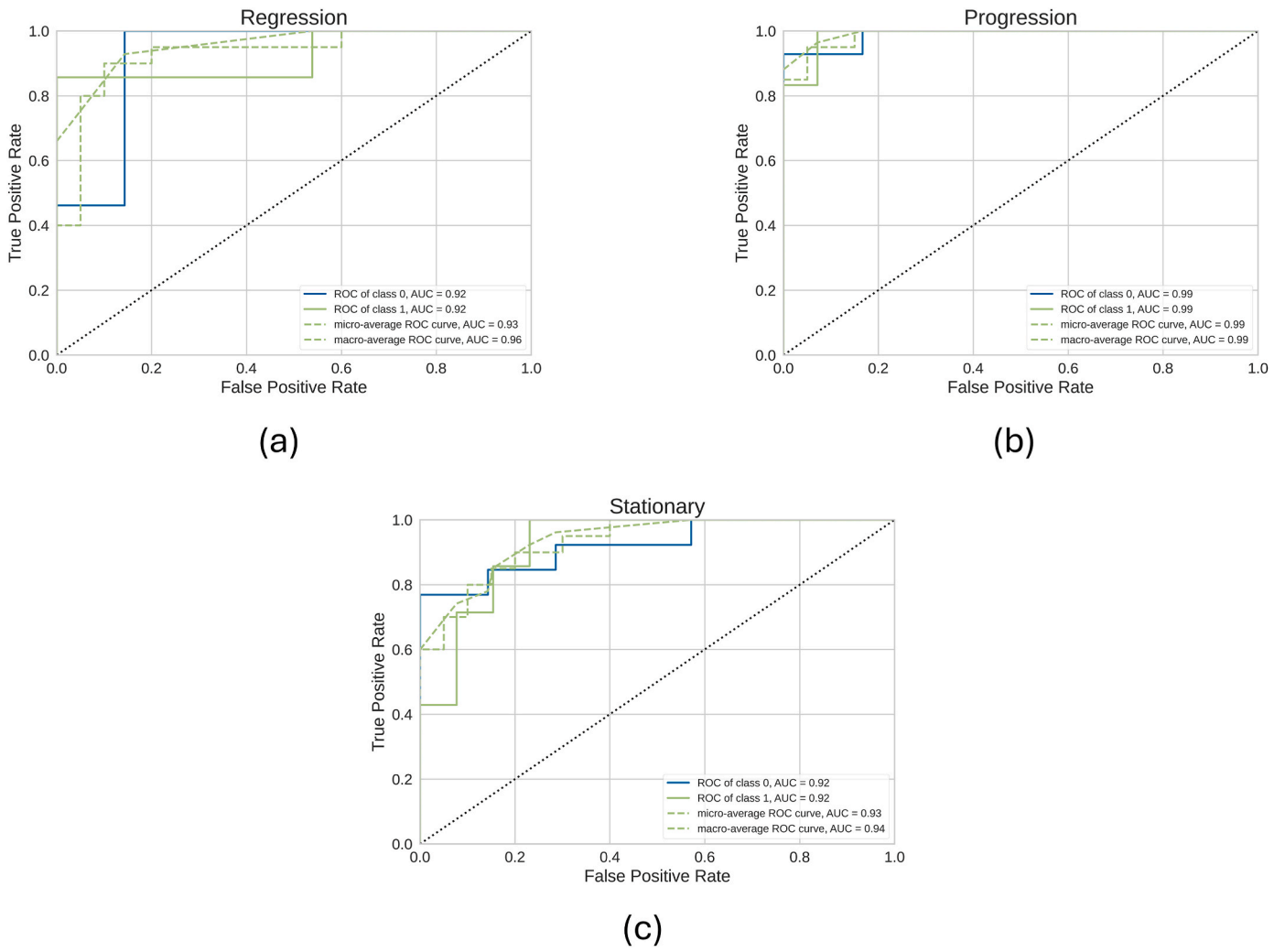


Fig. 3. ROC curves for the three classification tasks. The ROC curve of Class 0 represents the effectiveness in predicting the negative class, while the ROC curve of Class 1 shows the effectiveness in predicting the positive class. The micro-average curve represents the overall classification accuracy, where the true positive, false positive, true negative, and false negative values for each class are aggregated, and the curve is generated based on these combined values. The macro-average curve is the average of the individually calculated ROC curves for each class, balancing any potential class ratio differences.

4. Summary

Evaluating malignancy-mediated changes in serum N-glycosylation can be particularly significant in cancer research from both prognostic and diagnostic perspectives. Our research aimed to improve the reproducibility of the sample preparation method as well as to set up the most suitable ML-based data evaluation workflow to predict the effectiveness of chemotherapy treatments in lung cancer patients utilizing ultra-high resolution N-glycan profiling by capillary electrophoresis. Among the fine-tuned classifiers, the QDA classifier proved to be the most effective method, resulting in AUC values of 0.8290, 0.8295, and 0.8410 for the classes of disease regression, progression, and stationary, respectively. The results of other classifiers were erratic, but in all cases, they fell short of the QDA classifier. Although the average results of the QDA classifier were acceptably high, the trained classifier was able to achieve much higher AUC values for all three classification tasks using the same dataset (Fig. 2).

Our study also revealed that for different classification tasks, different N-glycan structures were crucial in predicting the efficacy of chemotherapy treatment for lung cancer patients. It is important to note that the examined dataset was small regarding a fully comprehensive data analysis perspective, therefore, further samples and studies are necessary to establish a stable and reliable approach. Consequently, our future goal is to continue refining the results, based on the N-glycome analysis of additional serum sample sets using the workflow presented in this paper.

CRediT authorship contribution statement

Agnes Vathy-Fogarassy: Writing – original draft, Validation, Supervision, Methodology, Conceptualization. **Veronika Gombas:** Writing – original draft, Visualization, Validation, Software, Methodology. **Rebeka Torok:** Data curation. **Gabor Jarvas:** Writing – original draft, Supervision, Data curation, Conceptualization. **Andras Guttman:** Writing – original draft, Supervision, Data curation, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT-4o to improve the language and readability of the paper. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the content of the published article.

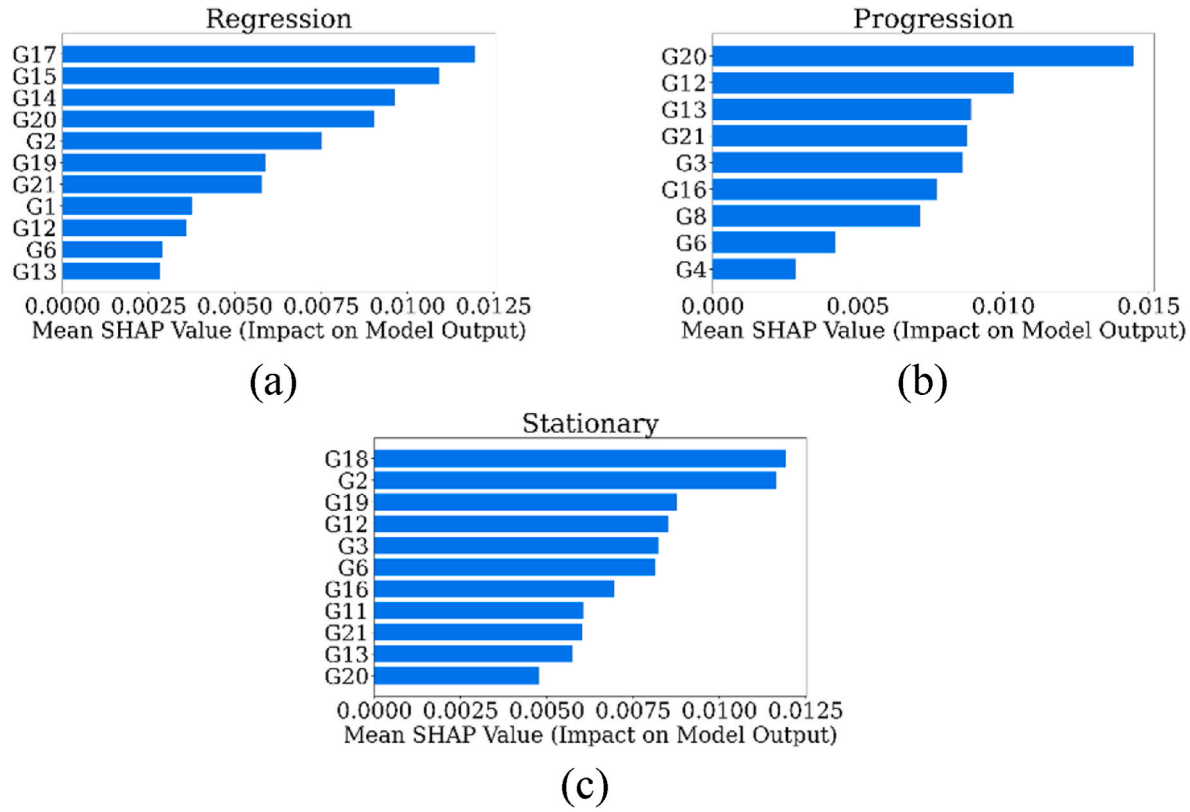


Fig. 4. The SHAP summary bar plot illustrates how individual N-glycans contribute to the accuracy of classification across different classification tasks: (a) Regression, (b) Progression, (c) Stationary. On the x-axis, the plot shows the average absolute SHAP values, which represent the extent of the contribution of each feature to the predictions of the models across 1000 different classification splits. The N-glycans on the y-axis are ordered by their impact on the predictions of the models.

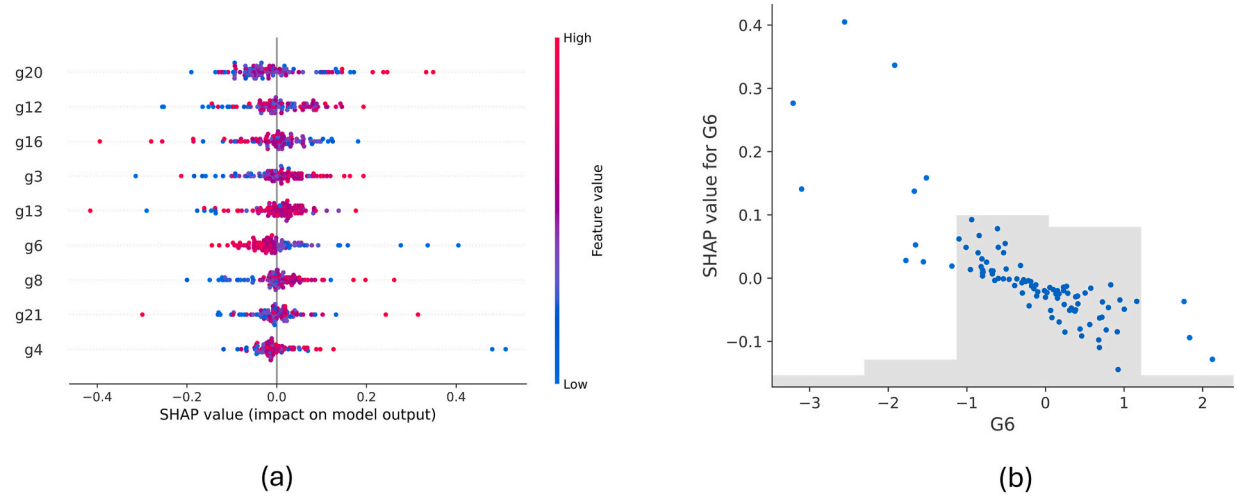


Fig. 5. (a) Beeswarm plot of the relevant N-glycans for classifying disease progression. On the y-axis, the N-glycans are ordered by the extent of their contribution to the predictions of the models. The x-axis displays the SHAP values. The color of each data point represents the magnitude of the value of the respective features. (b) SHAP dependence plot for N-glycan G6. The SHAP values for N-glycan G6 are depicted on the y-axis, while the x-axis represents the values for G6.

Funding sources

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 Cooperative Doctoral Program of the Ministry of Culture and Innovation, and by the University of Debrecen Program for Scientific Publication PTP2025.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This is contribution #222 of the Horváth Csaba Memorial Laboratory of Bioseparation Sciences. The authors gratefully acknowledge the valuable clinical support from Dr Eszter Csanky and Dr Miklos Szabo.

Appendix

Algorithm 1 Feature Selection and Feature Combination Detection

Method
Input: Feature set F
Output: Selected feature set F_{best}
Initialize the noise attribute set as $N = N_{new} = \varnothing$
Sort the list of features set F based on their relevance in descending order using the SFS method to create the list S.
Split the ordered list S at the midpoint of the list (rounded down to the nearest integer) into two parts: S_1 and S_2
Initialize the base set as $B \leftarrow S_1$
Initialize the set of features to be examined as $C \leftarrow S_2$
while $C \neq \varnothing$ **do**
Calculate the AUC value for the model built on attributes of B
for each subset C^* with size 1, 2, or 3 of C **do**
Form a new set by $B' \leftarrow B \cup C^*$
Calculate the AUC value for the model built on attributes of B'
end for
Select the feature set C^* for which the increase in AUC – when comparing the model built on B' to the one built on B – is the greatest, and – when C^* contains more than one element – this increase exceeds the sum of the individual AUC increases obtained by adding each element of C^* to B separately. For this C^*, define $R \leftarrow C^*$.
Identify the set of new noise attributes N_{new} as those attributes for which the AUC consistently decreases in all models where the attribute is included in C^*. Then, update N as $N \leftarrow N \cup N_{new}$.
Update the set B by $B \leftarrow R$
Update the set C by $C \leftarrow F - (B \cup N)$
end while
Return B as F_{best}

Data availability

The data that has been used is confidential.

References

[1] A. Varki, R.D. Cummings, J.D. Esko, H.H. Freeze, P. Stanley, C.R. Bertozzi, G. W. Hart, M.E. Etzler, Essentials of Glycobiology, second ed., Cold Spring Harbor Laboratory Press, New York, 2009.
[2] M. Hires, E. Jane, M. Mego, M. Chovanec, P. Kasak, J. Tkac, Glycan analysis as biomarkers for testicular cancer, Diagnostics 9 (2019) 156, https://doi.org/10.3390/diagnostics9040156.
[3] M. Hu, R. Zhang, J. Yang, C. Zhao, W. Liu, Y. Huang, H. Lyu, S. Xiao, D. Guo, C. Zhou, J. Tang, The role of N-glycosylation modification in the pathogenesis of liver cancer, Cell Death Dis. 14 (2023) 222, https://doi.org/10.1038/s41419-023-05733-z.
[4] B. Mészáros, G. Járvás, A. Farkas, M. Szigeti, Zs Kovács, R. Kun, M. Szabó, E. Csánky, A. Guttman, Comparative analysis of the human serum N-glycome in lung cancer, COPD and their comorbidity using capillary electrophoresis, J. Chromatogr. B 1137 (2020) 121913, https://doi.org/10.1016/j.jchromb.2019.121913.
[5] D. Thomas, A.K. Rathinavel, P. Radhakrishnan, Altered glycosylation incancer: a promising target for biomarkers and therapeutics, Biochim. Biophys. Acta Rev. Canc 1875 (2021) 188464, https://doi.org/10.1016/j.bbcan.2020.188464.
[6] P. Sitasuwan, T.W. Powers, T. Medwid, Y. Huang, B. Bare, L.A. Lee, Enhancing the multi-attribute method through an automated and high-throughput sample preparation, Mabs 13 (2021) 1978131, https://doi.org/10.1080/19420862.2021.1978131.
[7] K.W.P. Miller, N. Grossman, P. Haviernik, J. Wolff, C. Fu, B. Bare, E. Sindelar, A semi-automated tuberculosis testing workflow reduces manual hazardous sample handling and Hands-On time: a Proof-of-Concept study, SLAS Technol.: Translat. Life Sci. Innov. 25 (2020) 253–257, https://doi.org/10.1177/2472630319884519.
[8] Y. Wang, S. Lei, J. Dai, K. Yuan, Knowledge AI: new medical AI solution for medical image diagnosis, arXiv preprint arXiv:2101.03063 (2021).
[9] G.S. Mohammed, A comprehensive deep dive into machine learning: types, techniques, and unravelling its impact on diverse domains, Al-Salam J. Eng. Technol. 3 (2024) 24–37, https://doi.org/10.55145/ajest.2024.03.02.03.
[10] N. Papandrianos, E. Papageorgiou, Automatic diagnosis of coronary artery disease in SPECT myocardial perfusion imaging employing deep learning, Appl. Sci. 11 (2021) 6362, https://doi.org/10.3390/app11146362.
[11] S. Das, M.K. Dey, R. Devireddy, M.R. Gartia, Biomarkers in cancer detection, diagnosis, and prognosis, Sensors 24 (2024) 37, https://doi.org/10.3390/s24010037.
[12] S. Rajpal, A. Rajpal, A. Saggar, A.K. Vaid, V. Kumar, M. Agarwal, N. Kumar, Xai-methylmarker: explainable ai approach for biomarker discovery for breast cancer subtype classification using methylation data, Expert Syst. Appl. 225 (2023) 120130, https://doi.org/10.1016/j.eswa.2023.120130.
[13] D.B. Demirhan, H. Yılmaz, H. Erol, H.M. Kayili, B. Salih, Prediction of gastric cancer by machine learning integrated with mass spectrometry-based n-glycomics, Analyst 148 (2023) 2073–2080, https://doi.org/10.1039/D2AN02057B.
[14] B. Reider, M. Szigeti, A. Guttman, Evaporative fluorophore labeling of carbohydrates via reductive amination, Talanta 185 (2018) 365–369, https://doi.org/10.1016/j.talanta.2018.03.101.
[15] Cs Váradi, C. Lew, A. Guttman, Rapid magnetic bead based sample preparation for automated and high throughput N-glycan analysis of therapeutic antibodies, Anal. Chem. 86 (2014) 5682–5687, https://doi.org/10.1021/ac501573g.
[16] T. Jayashree, Shiva Prakash, K. Venugopal, Unsupervised feature extraction based on uncorrelated approach, Inf. Sci. 666 (2024) 120447, https://doi.org/10.1016/j.ins.2024.120447.
[17] Y. Wang, H. Tian, T. Li, X. Liu, A two-stage clonal selection algorithm for local feature selection on high-dimensional data, Inf. Sci. 677 (2024) 120867, https://doi.org/10.1016/j.ins.2024.120867.
[18] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput. Electr. Eng. 40 (1) (2014) 16–28, https://doi.org/10.1016/j.compeleceng.2013.11.024.
[19] M. Hearst, S. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines, IEEE Intell. Syst. Their Appl. 13 (4) (1998) 18–28, https://doi.org/10.1109/5254.708428.
[20] A. Tharwat, Linear vs. quadratic discriminant analysis classifier: a tutorial, International Journal of Applied Pattern Recognition 3 (2016) 145, https://doi.org/10.1504/IJAPR.2016.079050.
[21] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[22] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (2016) 785–794, https://doi.org/10.1145/2939672.2939785.
[23] B. Müller, J. Reinhardt, M.T. Strickland, Neural Networks: an Introduction, Springer Science & Business Media, 1995.
[24] S. Jasper, H. Larochelle, R.P. Adams, Practical bayesian optimization of machine learning algorithms, Adv. Neural Inf. Process. Syst. 25 (2012) 1206.2944, https://doi.org/10.48550/arXiv.1206.2944.
[25] B. Mészáros, G. Járvás, R. Kun, M. Szabó, E. Csánky, J. Abonyi, A. Guttman, Machine learning based analysis of human serum N-glycome alterations to follow

up lung tumor surgery, Cancers 12 (2020) 3700, https://doi.org/10.3390/cancers12123700.
[26] R. Torok, B. Meszaros, V. Gombas, A. Vathy-Fogarassy, M. Szabo, E. Csanky, G. Jarvas, A. Guttman, Predicting the effectiveness of chemotherapy treatment in lung cancer utilizing artificial intelligence-supported serum N-glycome analysis, Comput. Biol. Med. 186 (2025) 109681, https://doi.org/10.1016/j.compbiomed.2025.109681.
[27] S. Lundberg, S. Lee, A unified approach to interpreting model predictions, CoRR 1705 (2017) 07874, https://doi.org/10.48550/arXiv.1705.07874.