Inga Trulson*, Stefan Holdenrieder and Georg Hoffmann

Using machine learning techniques for exploration and classification of laboratory data

https://doi.org/10.1515/labmed-2024-0100
Received June 4, 2024; accepted June 26, 2024;
published online August 13, 2024

Abstract

Objectives: The study aims to acquaint readers with six widely used machine learning (ML) techniques (Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP), k-means, hierarchical clustering and the decision tree models (rpart and random forest)) that might be useful for the analysis of laboratory data.

Methods: Utilizing a recently validated data set from lung cancer diagnostics, we investigate how ML can support the search for a suitable tumor marker panel for the differentiation of small cell (SCLC) and non-small cell lung cancer (NSCLC).

Results: The ML techniques used here effectively helped to gain a quick overview of the data structures and provide initial answers to the clinical questions. Dimensionality reduction techniques such as PCA and UMAP offered insightful visualization and impression of the data structure, suggesting the existence of two tumor groups with a large overlap of largely inconspicuous values. This impression was confirmed by a cluster analysis with the k-means algorithm, indicative of unsupervised learning. For supervised learning, decision tree models like rpart or random forest demonstrated their utility in differential diagnosis of the two tumor types. The rpart model, which constructs binary decision trees based on the recursive partitioning algorithm, suggests a tree involving four serum tumor markers (STMs), which were confirmed by the random forest approach. Both highlighted pro-gastrin-releasing peptide (ProGRP), neuron specific enolase (NSE), cytokeratin-19 fragment (CYFRA 21-1) and cancer antigen (CA) 72-4 as key tumor markers, aligning with the outcomes of the initial statistical analysis. Cross-validation of the two proposals showed a higher area under the receiver operating characteristic (AUROC) curve of 0.95 with a 95 % confidence interval (CI) of 0.92–0.97 for the random forest model compared to an AUROC curve of 0.88 (95 % CI: 0.83–0.93).

Conclusions: ML can provide a useful overview of inherent medical data structures and distinguish significant from less pertinent features. While by no means replacing human medical and statistical expertise, ML can significantly accelerate the evaluation of medical data, supporting a more informed diagnostic dialogue between physicians and statisticians.

Keywords: k-means; machine learning; principal component analysis; random forest; decision tree

Introduction

Machine learning (ML) is increasingly becoming an integral part of laboratory medicine as evidenced by the rising number of published applications over the last decade [1–3]. Especially in clinical studies, where large quantities of laboratory values in combination with demographic, clinical and physiological data are collected, gaining a clear overview without deep statistical knowledge can be challenging.

In this study, we therefore tested the hypothesis whether integrating ML techniques into the standard statistical analysis process could help to better understand the data and thereby produce more medically meaningful results [4, 5]. This assumption has been supported by the fact that ML has now matured to the extent that standardized workflows are both possible and necessary, merging clinical expertise with innovative computational approaches [6].

A previously curated and statistically evaluated dataset was the source of our analysis [7]. It comprised patients diagnosed with various lung conditions, utilizing serum tumor markers (STMs) for differentiation at the time of diagnosis and before the initiation of therapy. A subset of this analysis focused on the application of STMs to distinguish small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). The importance of accurate lung cancer histology classification is underscored by its significant influence on treatment options and patient prognosis [8]. While traditional diagnostic methods are effective, they have limitations often requiring invasive procedures, highlighting the need

***Corresponding author: Inga Trulson**, Institute of Laboratory Medicine, German Heart Center Munich, Munich, Germany,
E-mail: trulson@dhm.mhn.de
Stefan Holdenrieder and Georg Hoffmann, Institute of Laboratory Medicine, German Heart Center Munich, Munich, Germany

for less invasive alternatives [9]. The investigated approach [7] sought to validate a minimal invasive, rapid alternative using a selected panel of protein markers in blood, potentially expediting the initiation of therapy.

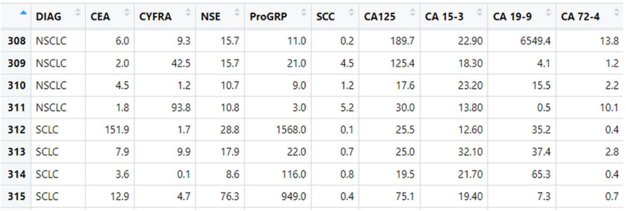
This study aims to evaluate how ML can be effectively applied to clinical data to reveal meaningful patterns that support medical diagnostics. We intend to conduct a 'machine learning exploratory data analysis' on this dataset with a focus on inherently explainable ML methods, to provide insights from this data that are not only accurate, but also comprehensible to clinicians. This approach could not only provide a detailed overview of the dataset but also help to understand the underlying structures within the data, thus setting the stage for more targeted and in-depth statistical analysis.

This exploratory analysis has not been designed to replace formal statistical evaluation but to enhance preliminary data exploration, thus facilitating the improved communication and collaboration between data scientists and statisticians. By detailing this workflow, we aim to contribute a practical analysis tool, that enhances early-stage data examination in a comprehensible way and supports effective decision-making in a clinical research setting.

Materials and methods

Study population and data collection

Detailed information on the original study population as well as on preanalytical and analytical methods can be obtained from a previously published study [7]. Briefly, blood samples were collected from 436 patients with non-small cell lung cancer and small cell lung cancer at the time of diagnosis and before the initiation of treatment. The values of nine STMs, CYFRA 21-1 (cytokeratin-19 fragment), CEA (carcinoembryonic antigen), NSE (neuron specific enolase), ProGRP (pro-gastrin-releasing peptide), SCC (squamous cell carcinoma antigen) and carbohydrate markers CA 125, CA 15-3, CA 19-9 and CA 72-4 were assessed (Figure 1).



	DIAG	CEA	CYFRA	NSE	ProGRP	SCC	CA125	CA 15-3	CA 19-9	CA 72-4
308	NSCLC	6.0	9.3	15.7	11.0	0.2	189.7	22.90	6549.4	13.8
309	NSCLC	2.0	42.5	15.7	21.0	4.5	125.4	18.30	4.1	1.2
310	NSCLC	4.5	1.2	10.7	9.0	1.2	17.6	23.20	15.5	2.2
311	NSCLC	1.8	93.8	10.8	3.0	5.2	30.0	13.80	0.5	10.1
312	SCLC	151.9	1.7	28.8	1568.0	0.1	25.5	12.60	35.2	0.4
313	SCLC	7.9	9.9	17.9	22.0	0.7	25.0	32.10	37.4	2.8
314	SCLC	3.6	0.1	8.6	116.0	0.8	19.5	21.70	65.3	0.4
315	SCLC	12.9	4.7	76.3	949.0	0.4	75.1	19.40	7.3	0.7

Figure 1: Excerpt from the original data set.

The evaluation study was conducted in accordance with the ethical standards set out in the Declaration of Helsinki, and the use of anonymous test data for statistical evaluation purposes was approved by the Ethics Committee of the Ludwig-Maximilians-Universität (LMU) in Munich (UE-Nr 114-13), as specified in the Institute Review Board approval.

Statistical analysis

Differences in median STM levels between the two different cancer classes were evaluated using the Mann–Whitney U test (data not shown). p-values <0.05 were considered statistically significant. The distributions, pairwise relationships, and correlation coefficients within the dataset were examined and depicted in a scatterplot matrix using the ggpairs function from the GGally package in R.

When a model required scaled values, a robust scaling method [10] was applied to standardize the range of values across different variables, ensuring they are all on a comparable scale as follows:

$$x_{\text{scaled}} = \frac{\text{x} - \text{Median}\,(\text{X})}{\text{IQR}\,(\text{X})}$$

where x is the original value, X is the vector of all values x and IQR is the interquartile range.

To account for the right-skewed distribution of the markers, log-values of the markers were used. Missing values were imputed with the median of all patients with NSCLC and SCLC.

Six different widely used ML tools were used to analyze the data. Dimensionality reduction was performed using Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP). For unsupervised learning the k-means algorithm was employed and further a hierarchical clustering analysis was conducted and visualized in a heatmap. For supervised learning two decision tree (DT) models were explored: Recursive Partitioning and Regression Trees (rpart) and Random Forest analysis. The validation of tree-based algorithms was conducted by leave-one-out cross-validation to avoid overfitting, and accuracy was calculated. As a performance metric, the area under the Receiver Operating Characteristic (AUROC) curves for the DT models were calculated, to encounter for the imbalance of the dataset, using the pROC package [11]. The confidence interval for the AUROC analysis was computed based on 5,000 bootstrap samples. SCLC was designated as the positive class.

Statistical analysis and ML modelling were performed using R software (version 4.3.2). The ML models used for the visualization and data analysis are depicted and summarized in Table 1. The Table also includes the corresponding

Table 1: Algorithms and models used for data analysis. The corresponding code includes the respective R functions followed by their parameters in parenthesis, where x stands for the data frame to be analyzed, DIAG stands for the diagnosis code (NSCLC, SCLC), centers stand for the number of clusters to find, type stands for a certain type of plot, where the split variable names a depicted in the nodes of the tree.

Algorithm/model	Aim/purpose	Packages	Code

PCA	Dimensionality reduction+visualisation	stats (base R)	pca <- prcomp(x)
			plot(pca)

UMAP	Dimensionality reduction+visualisation	umap [27]	umap <- umap(x$layout)
			plot(umap)

k-means	Clustering
	Visualisation
		stats (base R)
		factoextra [28]
			km <- kmeans(x, centers=3)
			fviz_cluster(km, data=x)

Hierarchical clustering	Clustering+visualisation	gplots [29]	heatmap.2(x)

rpart	Classification
	visualisation
		rpart [30]
		rpart.plot [31]
			rp <- rpart(formula=DIAG ~., data=x)
			rpart.plot(rp, type=5)

Random forest	Classification	Tree [32]
		randomForest [33]
			rf <- randomForest(formula=DIAG ~., data=x)
			print(rf)

Explorative data analysis	generation of scatterplot matrix	GGally [34]	ggpairs(x)

PCA, Principal Component Analysis; UMAP, Uniform Manifold Approximation and Projection; rpart, Recursive Partitioning and Regression Trees.

code and the relevant R packages and functions used for each model.

Results

Composition of the cohorts

The investigated patient cohort included 308 patients with non-small cell lung cancer (98 female; 210 male) and 128 patients with small cell lung cancer (35 female; 93 male). There were no significant differences between the groups in terms of age or sex. For more detailed information see [7].

Data preparation

Due to the right-skewed distribution of the markers (see Figure 2) log-transformed values were used for further processing of the data. To ensure the comparability of different levels of marker values, robust z-score scaling was applied. This method provided a more balanced distribution compared to the traditional z-score scaling technique (see Figure 3).

Missing values of serum tumor markers were imputed in eight cases for NSCLC and one case for SCLC. No noteworthy variations in the outcomes were evident when cases with missing values were omitted from the analysis.

Exploratory data analysis

The aim of exploratory data analysis (EDA) is to gain an understanding of the structures and meanings of the experimental data, to better assess their suitability for the research question of the study (here differentiation between SCLC and NSCLC). Figure 4 illustrates a typical example of such an analysis which, however, clearly shows that multivariate data sets can only be insufficiently characterized by a series of 2-dimensional scatterplots.

The distributions, relationships, and correlations among the logarithms of the investigated marker concentrations are summarized in Figure 4.

The highest correlation coefficients are observed between ProGRP and NSE (0.538, p<0.001) as well as CA 15-3 and CA 125 (0.517, p<0.001), indicating similarities in their clinical significance. Some of the correlation coefficients are negative, e.g. ProGRP vs. CA 15-3 or CA 72-4, which results in scatterplots that indicate the existence of two cohorts.

Machine learning data analysis and application

Dimensionality reduction and visualization

To visualize higher-dimensional data on the one hand and to reduce the complexity of such data for ML on the other, a PCA and UMAP analysis was performed (Figure 5). Both algorithms use statistical methods to reduce the number of variables (in this case tumor marker concentrations) by combining them into new composite variables. While PCA is a more than 100 years old technique, UMAP was developed in 2018 by McInnes et al. [12] and produces more informative graphics by better preserving the original distances between the data points.

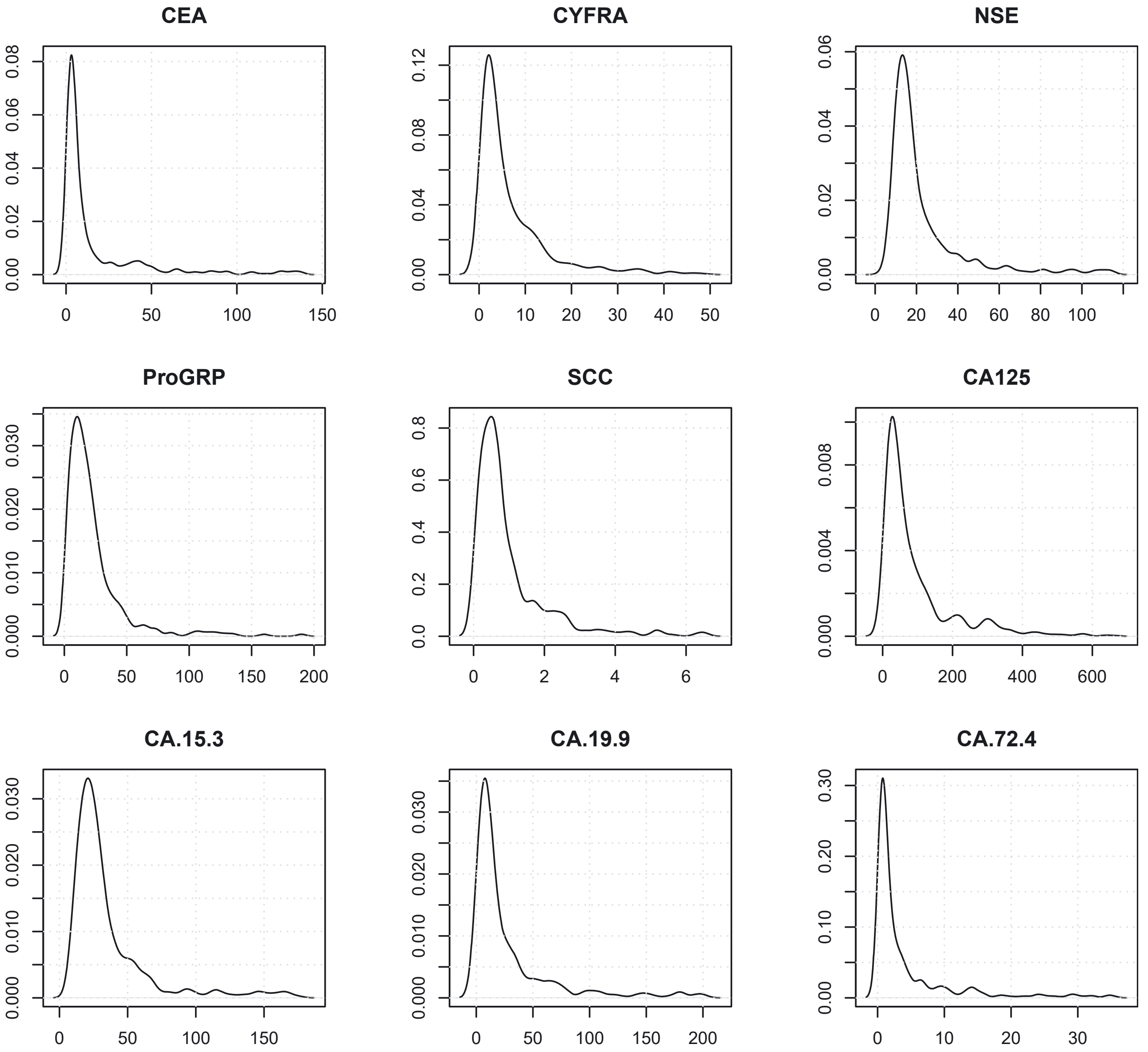


Figure 2: Right-skewed distribution of the investigated tumor markers. CEA, carcinoembryonic antigen; CYFRA, cytokeratin-19 fragment; NSE, neuron specific enolase; ProGRP, pro-gastrin-releasing peptide; SCC, squamous cell cancer antigen; CA, cancer antigen.

The PCA plot shows a region of high point density, indicating concentrations of observations with similar characteristics within the reduced two-dimensional space. In addition, there are two less dense “arms”, suggesting a possible bifurcation in the data structure, which can be attributed to the two tumor types, when the data points are labelled (Figure 5, top right). The principal components, especially the first principal component (PC1), summarize the variance within the data with ProGRP and NSE being the most influential biomarkers. This influence is quantified by their high loadings in the PCA, indicating that these biomarkers significantly contribute to the variance captured by the PCA1 (data not shown).

The UMAP plot keeps the individual data points better apart, supporting more clearly the impression of two distinguishable cohorts with a broad overlap. The labelled datapoints (Figure 5, bottom right) clearly suggest, that a group of patients with SCLC can distinctly be separated by the investigated biomarkers (variables).

To identify and separate potential patient subgroups even better, we applied the k-means cluster algorithm as one

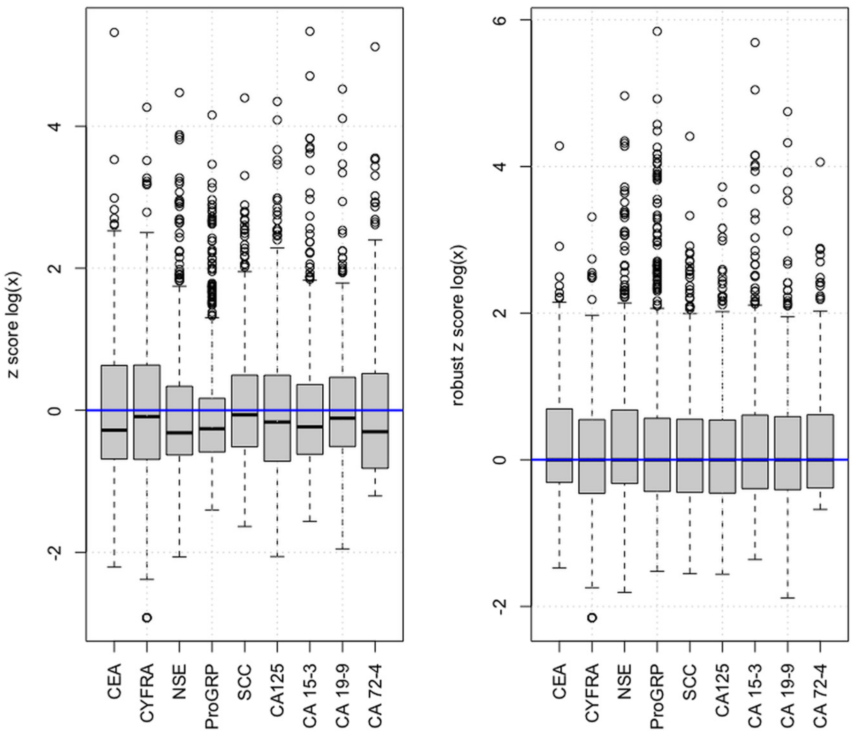


Figure 3: Comparison of z-score scaling (left) and robust z-score scaling (right).

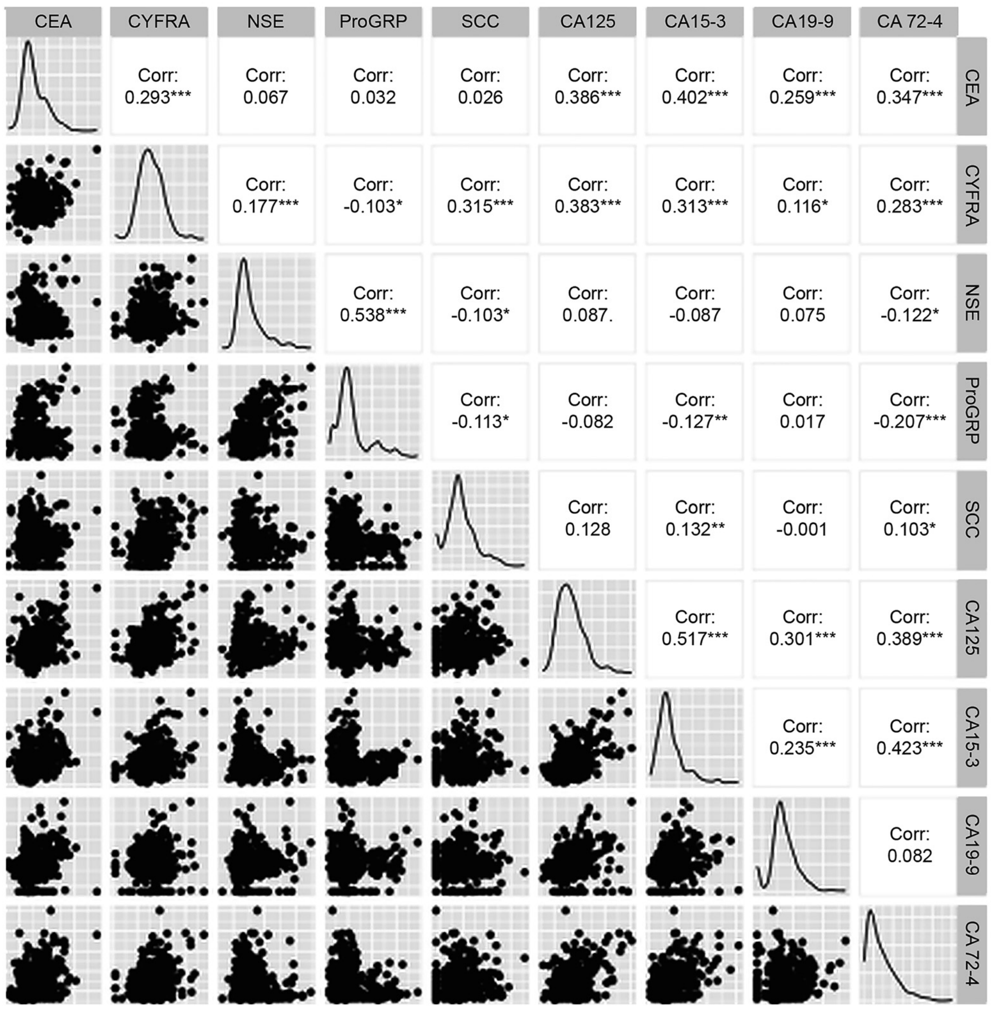


Figure 4: Scatterplot showing the distribution and correlation among the investigated markers in all lung cancer patients.

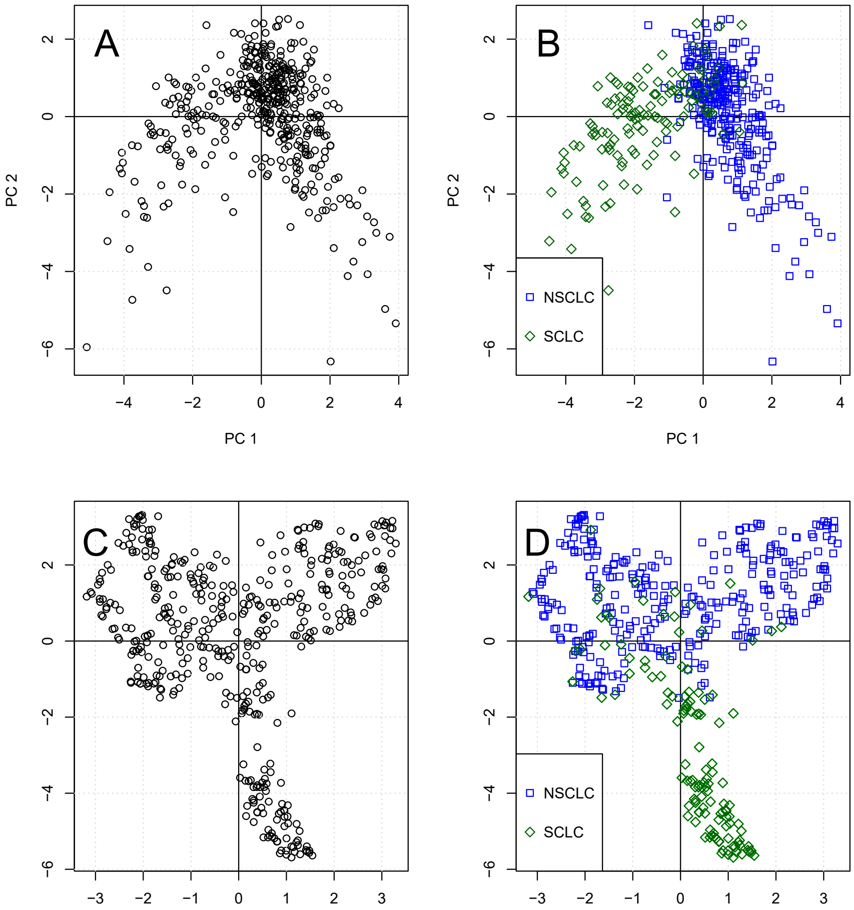


Figure 5: Dimensionality reduction analysis. Principal Component Analysis (PCA) (A), labelled datapoints in the PCA (B), Uniform Manifold Approximation and Projection (UMAP) (C), labelled datapoints in the UMAP analysis (D). Unsupervised learning – k-means and hierarchical clustering.

of two examples of unsupervised ML. This technique revealed three clusters within the PCA-reduced two-dimensional space (Figure 6): cluster 1 comprising 248 patients, displayed a mix of histologies with values mainly below the cut-off values specified by the manufacturers. Cluster 2, with 83 patients, predominantly consisted of SCLC cases, while cluster 3 included 105 patients, mostly with NSCLC histology (Figure 6). The optimal number of

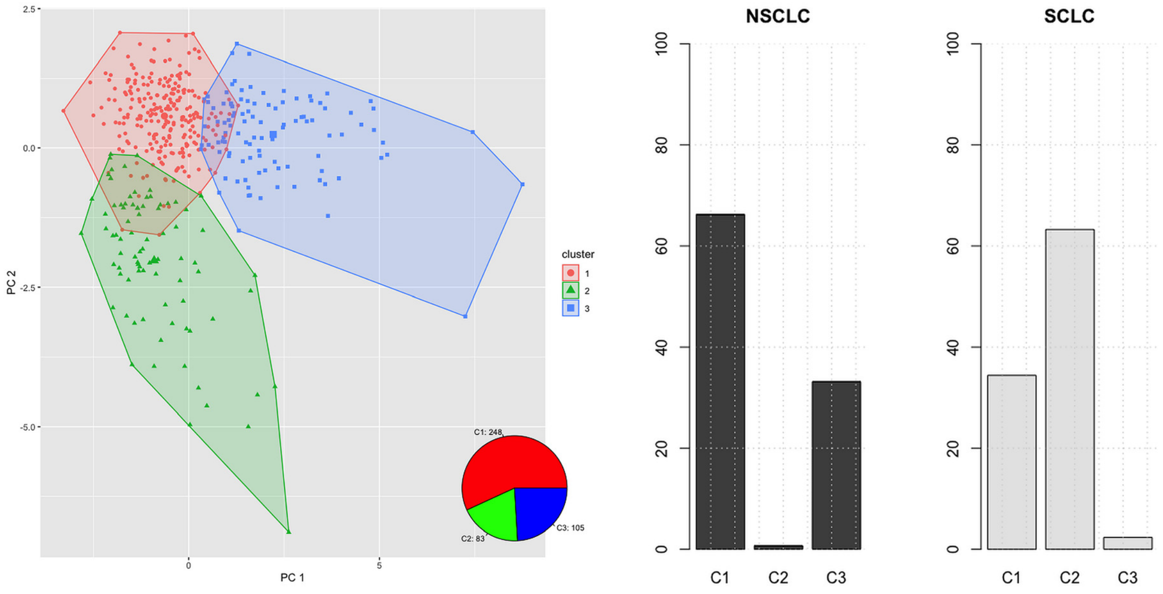


Figure 6: The k-means analysis reveals three clusters (left): C1 (red), C2 (green) and C3 (blue). C1 is composed of mixed histology, whereas C2 is mainly composed of small cell lung cancer (SCLC) patients and C3 of non-small cell lung cancer (NSCLC) patients (right).

three clusters was determined by calculating the total within-cluster sum of squares and analyzing the elbow plot visually [13].

Figure 7 shows that the three clusters identified by k-means make sense in terms of tumor marker profiles: ProGRP and NSE are increased in cluster 2, whereas CYFRA 21-1, CEA and carbohydrate markers are elevated in cluster 3. Cluster 1 appears as a mixture of both tumor types with more or less inconspicuous tumor marker values at the level of the respective decision limits (dashed lines in Figure 7).

As a second example of unsupervised ML, we performed hierarchical clustering and visualized the results in a heatmap using the heatmap.2 function from the gplots package (Figure 8, Table 1). The dendrograms alongside the heatmap show the hierarchical structure of the data, with both rows and columns being clustered based on similarity [14]. The black rectangles in Figure 8 highlight three specific clusters: two of them (A and B) represent cases with high ProGRP and/or NSE levels, and a third cluster with high values of CYFRA 21-1, CEA and CA markers. Additionally, there is also a

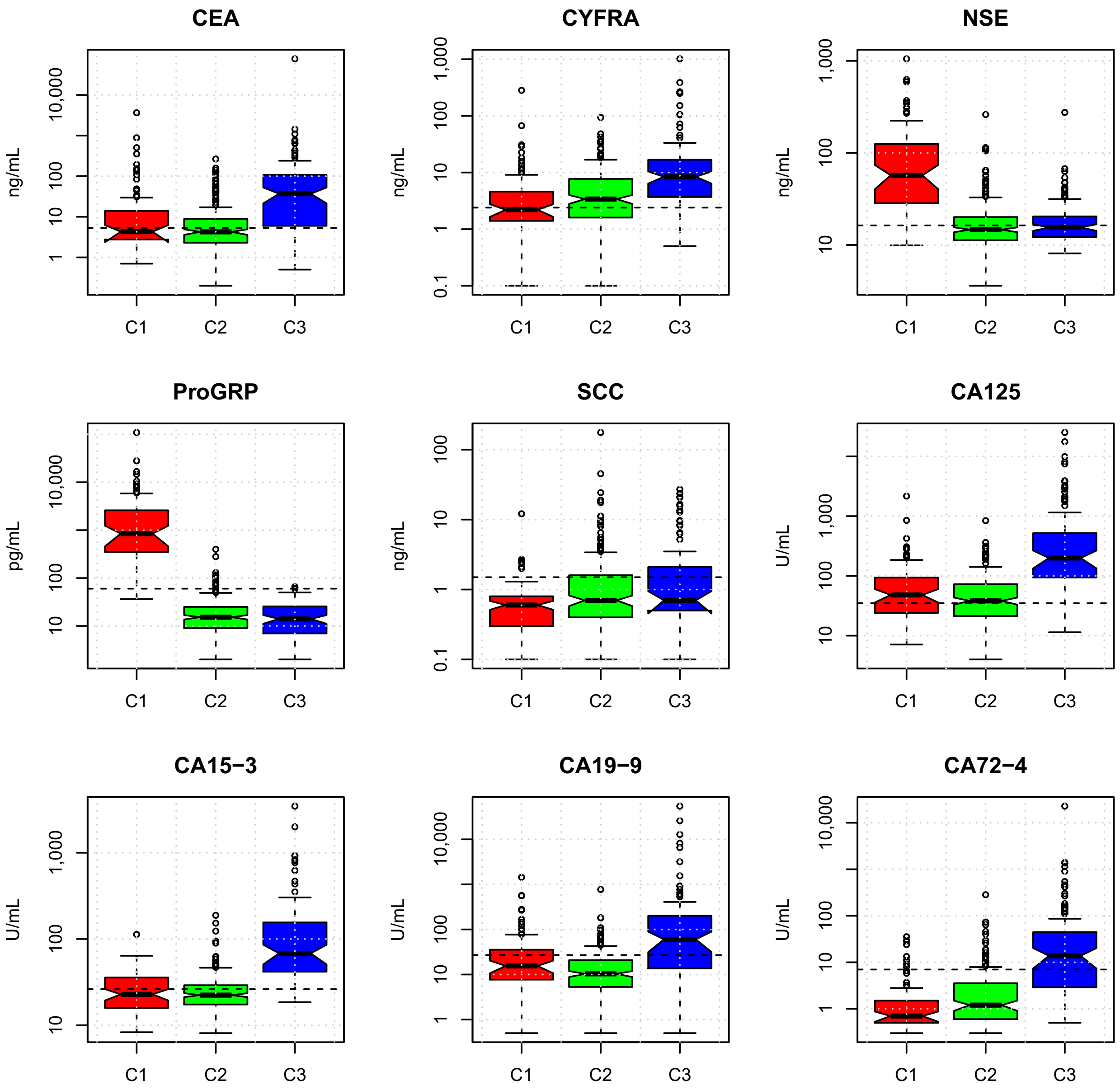


Figure 7: Tumor marker levels within the different clusters of the k-means analysis. Dashed lines indicate the cut-off values for the respective tumor marker specified by the manufacturers (95th percentile of healthy controls).

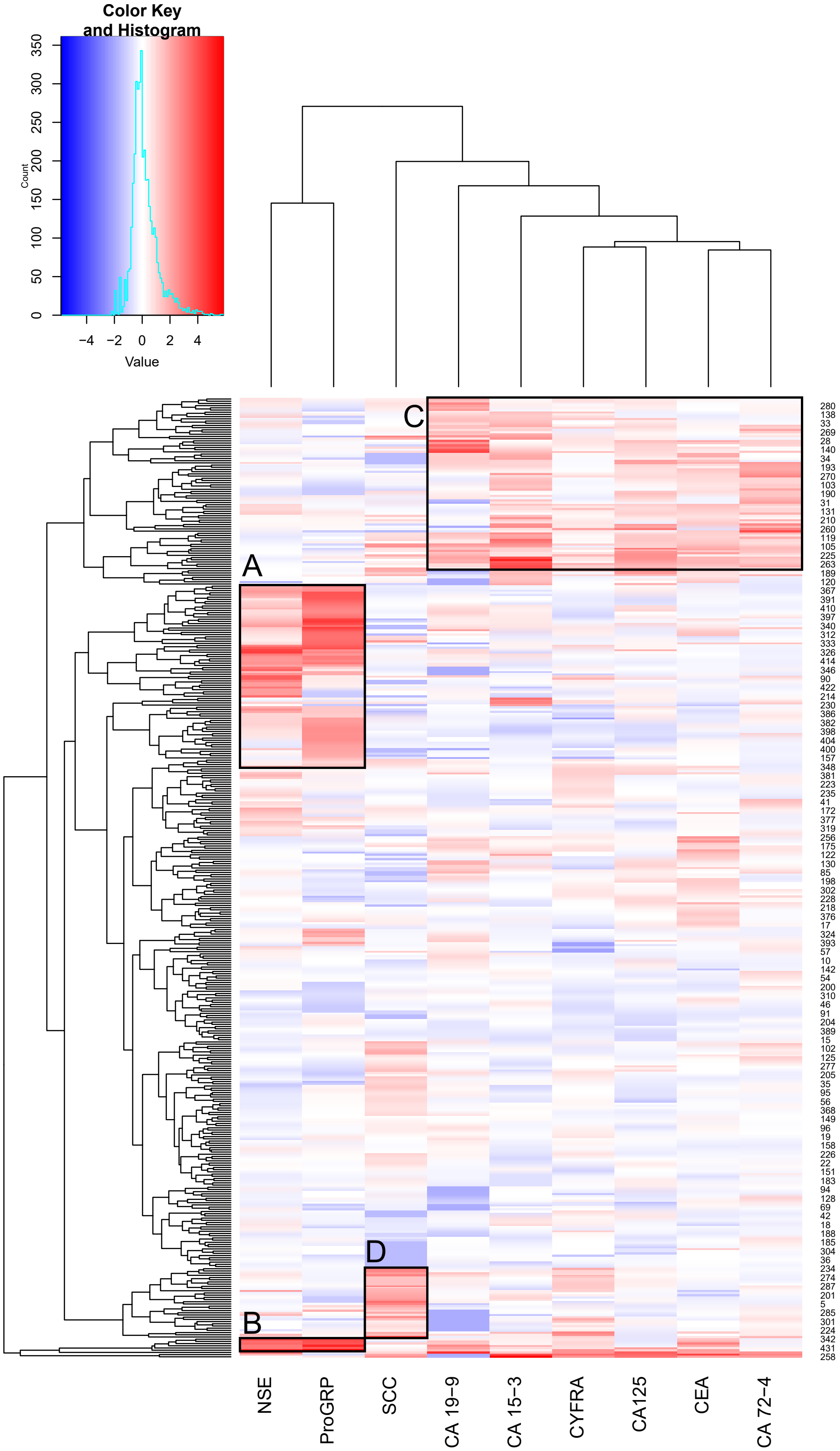


Figure 8: Hierarchical clustering visualized by a heatmap with dendrograms. Specific clusters (A–D) are framed with black rectangles. Supervised learning – tree-based models.

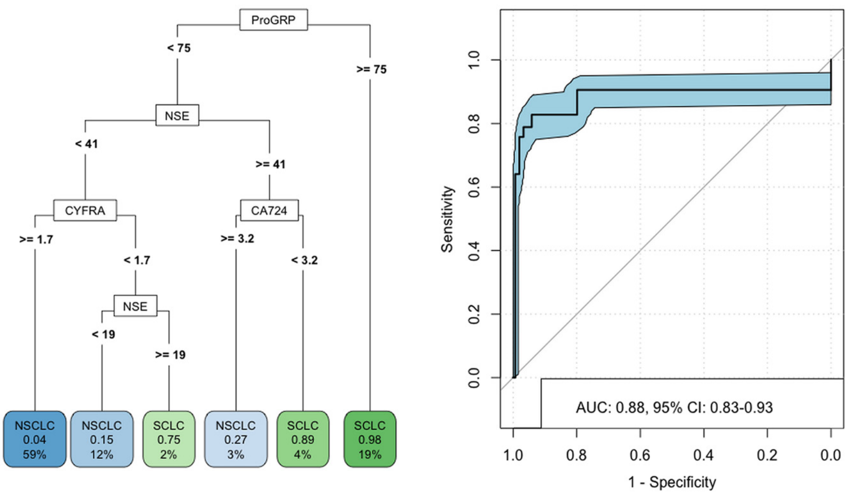


Figure 9: Decision tree with the rpart model (left), area under the receiver operating characteristic curve (AUC) and corresponding 95 % confidence interval (CI) for cross-validated data (right).

smaller group with high SCC values (D), which may represent a histological subgroup.

Since the supervised learning models investigated here do not require any scaling, we applied the respective algorithms to the original values (Figure 1). The rpart algorithm employs recursive binary partitioning to minimize the error in prediction and generates a quite detailed decision tree as depicted in Figure 9. Notably, high ProGRP and NSE levels are identified by the model as key predictors for the diagnosis of SCLC. The validation with leave-one-out cross-validation provides an area under the receiver operating characteristic (AUROC) curve of 0.88 with a 95 % confidence interval (CI) of 0.83–0.93 (Figure 8) and an accuracy of 92 %.

Further, we applied a random forest algorithm, generating a large set of 500 decision trees, to reduce overfitting. Each tree was constructed from bootstrapped samples drawn from the dataset, as described by Breimann (2001) [15]. As a visual presentation of each features' significance influencing the model's predictive accuracy, a 'Variable Importance Plot' is shown in Figure 10. The plot depicts the Mean Decrease Gini values, where higher values indicate higher significance for partitioning the data in the defined classes. Notably, ProGRP, NSE and CYFRA 21-1 emerged as the most influential features with the highest Mean Decrease Gini values. The AUCROC curve of the cross validated data was 0.95 with a 95 % CI of

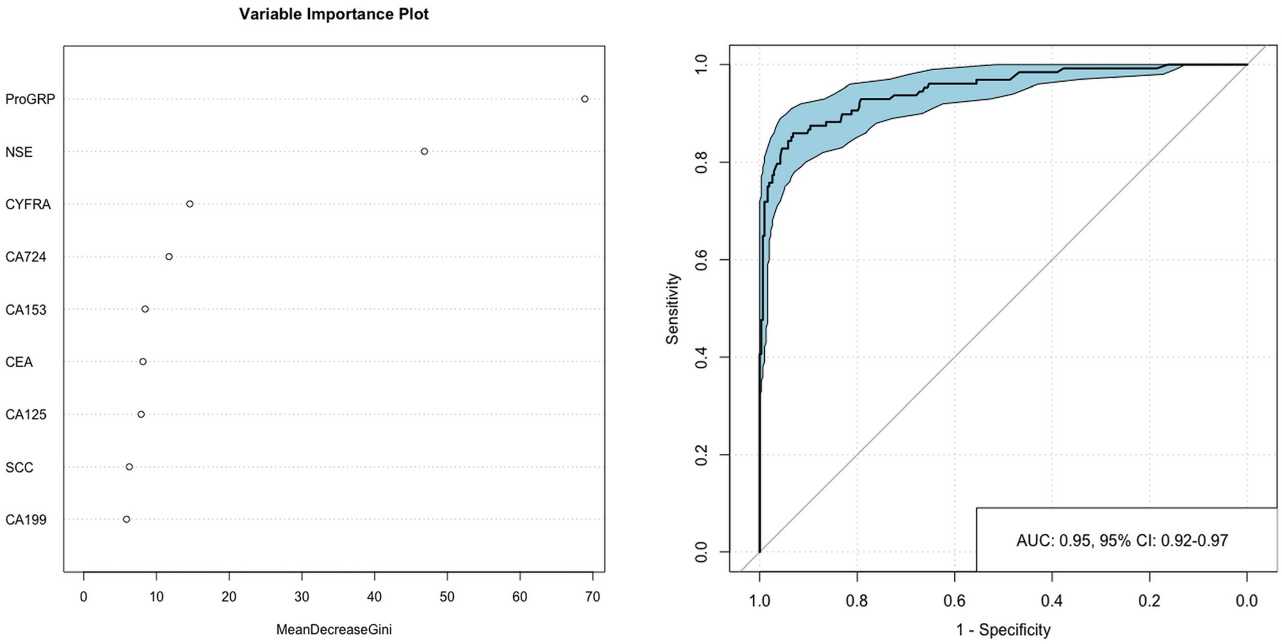


Figure 10: Variable importance plot from the random forest analysis (left) and receiver operating curve with corresponding area under the curve (AUC) and 95 % confidence interval (95 % CI) for cross-validated data.

0.92–0.97 (see Figure 10) and a calculated accuracy in the leave-one-out cross validation of 91 %.

Discussion

After completing several basic statistical analysis and exploration of the data, we propose a three-step ML workflow that could be applied to similar datasets as shown in Figure 1. However, it is important to emphasize that the choice of ML techniques and workflows depend to a large extent on the specific research question, as well as the characteristics and complexity of the dataset. These factors greatly influence the suitability and effectiveness of the different algorithms.

As a first step for the two-dimensional visualization of the data, we applied dimensionality reduction with a classical (PCA) and even more effectively with a newer (UMAP) method. This provided a plausible insight into the underlying multidimensional data structure. The graphical presentation suggested the presence of distinct clusters within the dataset, with UMAP providing a more distinguishable subgroup. NSE and ProGRP were identified as the biomarkers with the greatest impact on variance within the data in the PCA. Notably, the substantial overlap of initially appearing inconspicuous values may also encompass tumors with mixed histologies or potentially distinct subhistologies of NSCLC.

This impression was further confirmed by the k-means clustering algorithm (unsupervised learning), in the second step of our workflow. The clustering analysis identified groupings that align with our initial observations from the dimensionality reduction methods. In unsupervised ML, models analyze unlabeled datasets and are trying to identify patterns of relationships between variables [2], which reflect the greatest possible variance within the data. Although our data set for this unsupervised analysis was intentionally unlabeled, we applied labels posthoc – in our case the diagnoses of the patients. By color coding the data points according to their labels, we found that our clustering model effectively separated the two diagnostic groups. The second cluster consisted almost exclusively of SCLC patients, again with elevated NSE and ProGRP levels (Figure 7). The hierarchical clustering corroborates the findings of the k-means analysis; however, its interpretation may be more challenging and is prone to error for non-experts [14].

The application of ML in laboratory medicine is still limited [16, 17], often challenged by low model transparency [18] and insufficient interpretability of the generated results [4, 16, 19]. To encounter this issue, explainable AI (*XAI*) techniques have been developed [16, 20], that are inherently interpretable, always in consideration of balancing the performance of the model used and the interpretability of the results [3, 16, 21]. This is an effective strategy enhancing trust and understanding among medical practitioners. Therefore, as a final step, we generated two decision tree-based models for the differential diagnosis of SCLC and NSCLC, as they provide insights that are self-explanatory for medical experts.

In our study, the rpart model suggests a tree involving four tumor markers: high ProGRP or NSE levels would suggest SCLC histology and elevated CYFRA and CA 72-4 levels in absence of high ProGRP and NSE levels NSCLC histology. To avoid model overfitting, we performed a random forest analysis, which emphasized the same four markers, indicating their role in predicting outcomes effectively while maintaining model generalizability across various datasets. Cross-validation of the two proposals with the leave-one-out method showed a higher AUC of 0.95 (95 % CI: 0.92–0.97) for the random forest model, however slightly more accurate results for rpart (92 % vs. 91 %, respectively).

These findings are consistent with those of a much more sophisticated statistical analysis reported in the original study. Based on multivariate regression models, ProGRP and NSE emerged as the most effective marker combination in distinguishing SCLC from NSCLC (AUC 0.90, 95 % CI: 0.86–0.94). Adding CYFRA 21-1 in a triple combination increased the AUC to 0.93 (95 % CI: 0.91–0.96), whereas adding CA72-4 as the best four marker combination (AUC 0.94, 95 % CI: 0.92–0.97) yielded just a marginal improvement. These markers were also particularly influential in the PCA and were included in the DT model analysis, providing a quite good categorization.

In the present study with an admittedly simple data set, the quick and easy analysis with ML models provided very similar results regarding biomarker selection for the histological distinction of lung cancer. This not only gave us a first impression of the data but could also point the way for future ML applications in biomarker studies.

Although our primary objective was not the evaluation of different classification models, it is worth mentioning that especially supervised ML models are eligible for assessing medical data. Tree-based models in particular are common in this field [2, 3]. Hoffmann et al. [22] support the use of tree models as a valid auxiliary tool for the validation of medical expert opinions. Vogg et al. [23] recently proposed a simple decision tree using the ctree function to distinguish between adrenocortical carcinoma and adrenocortical adenoma based on two urinary steroid markers. By categorizing patients into three risk levels, the rate of unnecessary surgical procedures could potentially be reduced. Steinbach et al. [24] applied ML techniques to predict sepsis at the time of intensive care unit admission using complete blood count values and clinical data, which is quick and widely available.

These strategies are examples of how ML can use easily accessible data to provide prompt, actionable insights, which possibly enhance patient care in time-sensitive situations. However, it must be emphasized that before delving into ML modeling, it is imperative to ensure that the data on which the model is trained is clear, pertinent, and organized in a way that optimizes the performance of ML algorithms. Data preprocessing, which is the largest time-consuming phase in ML [25], is vital because the quality of the data used in the model, significantly influences the accuracy and reliability of the results [26]. Consequently, once the data is well-prepared, the actual ML modelling process often requires only a few lines of code and hence can easily be integrated into the normal preprocessing workflow.

Conclusions

Machine learning does not reach the depth of detail achieved by statistical experts, for example regarding the discriminatory power of elaborated biomarker combinations, but it does provide a good overview of the inherent medical data structures and easily distinguishes relevant from less meaningful biomarker data. While, by no means, replacing human expertise, ML can significantly speed up the evaluation of medical studies and support the dialogue between physicians and statisticians.

Research ethics: Research involving human subjects complied with all relevant national regulations, institutional policies and is in accordance with the tenets of the Helsinki Declaration and has been approved by the Ethics Committee of the Ludwig-Maximilians-Universität (LMU) in Munich (UE-Nr 114-13).
Informed consent: Not applicable.
Author contributions: The authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: The authors state no conflict of interest.
Research funding: None declared.
Data availability: The raw data can be obtained on request from the corresponding author.

References

1. Cabitza F, Banfi G. Machine learning in laboratory medicine: waiting for the flood? Clin Chem Lab Med 2018;56:516–24.
2. Rabbani N, Kim G, Suarez C, Chen J. Applications of machine learning in routine laboratory medicine: current state and future directions. Clin Biochem 2022;103:1–7.
3. Cubukcu H, Topcu D, Yenice S. Machine learning-based clinical decision support using laboratory data. Clin Chem Lab Med 2024;62:793–823.
4. Mao L, Wang H, Hu LS, Tran NL, Canoll PD, Swanson KR, et al Knowledge-informed machine learning for cancer diagnosis and prognosis: a review. 2024. https://doi.org/10.48550/arXiv.2401.06406.
5. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2014;13:8–17.
6. Master S, Badrick T, Bietenbeck A, Haymond S. Machine learning in laboratory medicine: recommendations of the IFCC working group. Clin Chem 2023;69:690–8.
7. Trulson I, Klawonn F, von Pawel H, Holdenrieder S. Improvement of differential diagnosis of lung cancer by use of multiple protein tumor marker combinations. Tumor Biol 2024;46:81–98.
8. Ciofiac CM, Mămuleanu M, Florescu LM, Gheonea IA. CT imaging patterns in major histological types of lung cancer. Life 2024;14:462.
9. Holdenrieder S. Biomarkers along the continuum of care in lung cancer. Scand J Clin Lab Invest Suppl 2016;245:S40–5.
10. Berthold MR, Borgelt C, Höppner F, Klawonn F, Silipo R. Data understanding. In: Guide to intelligent data science. Texts in computer science, 2nd ed. Cham: Springer; 2020:127–56 pp.
11. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sachez J-C, et al. pROC: display and analyze ROC curves. https://CRAN.R-project.org/package=pROC [Accessed 4 May 2024].
12. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for dimension reduction 2018. https://doi.org/10.48550/arXiv.1802.03426 [Accessed 4 May 2024].
13. Yuan C, Yang H. Research on K-value selection method of K-means clustering algorithm. Multidiscip. Res. J. 2019;2:226–35.
14. Engle S, Whalen S, Joshi A, Pollard KS. Unboxing cluster heatmaps. BMC Bioinf 2017;18:63.
15. Breiman L. Random forests. Mach Learn 2001;45:5–32.
16. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform 2021;113:103655.
17. Peterson ED. Machine learning, predictive analytics, and clinical practice: can the past inform the present? JAMA 2019;322:2283–4.
18. Harrison JH, Gilbertson JR, Hanna MG, Olson NH, Seheult JN, Sorace JM, et al. Introduction to artificial intelligence and machine learning for pathology. Arch Pathol Lab Med 2021;145:1228–54.
19. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44–56.
20. Minh D, Wang HX, Li YF, Nguyen TN. Explainable artificial intelligence: a comprehensive review. Artif Intell Rev 2022;55:3503–68.
21. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. Entropy 2020;23:18.
22. Hoffmann G, Bietenbeck A, Lichtinghagen R, Klawonn F. Using machine learning techniques to generate laboratory diagnostic pathways-a case study. J Lab Precis Med 2018;3:58.
23. Vogg N, Müller T, Floren A, Dandekar T, Riester A, Dischinger U, et al. Simplified urinary steroid profiling by LC-MS as diagnostic tool for malignancy in adrenocortical tumors. Clin Chim Acta 2023;543:117301.
24. Steinbach D, Ahrens PC, Schmidt M, Federbusch M, Heuft L, Lubbert C, et al. Applying machine learning to blood count data predicts sepsis with ICU admission. Clin Chem 2024;70:506–15.
25. Tawalkuli A, Havers B, Gulisano VM, Kaiser D, Engel T. Survey: time-series data preprocessing: a survey and an empirical analysis. J Eng Res 2024. https://doi.org/10.1016/j.jer.2024.02.018.

26. Albahra S, Gorbett T, Robertson S, D’Aleo G, Ockunzzi S, Lallo D, et al. Artificial intelligence and machine learning overview in pathology & laboratory medicine: a general review of data preprocessing and basic supervised concepts. Semin Diagn Pathol 2023;40:71–87.
27. Konopka T. Umap: Uniform Manifold Approximation and Projection; 2023. https://CRAN.R-project.org/package=umap [Accessed 4 May 2024].
28. Kassambara A, Mundt F. Factoextra: extract and visualize the result of multivariate data analyses; 2020. https://CRAN.R-project.org/package=factoextra [Accessed 4 May 2024].
29. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, et al. gplots: various R programming tools for plotting data; 2024. https://CRAN.R-project.org/package=gplots [Accessed 4 May 2024].
30. Therneau T, Atkinson B. Rpart: recursive partitioning and regression trees; 2023. https://CRAN.R-project.org/package=rpart [Accessed 4 May 2024].
31. Milborrow S. Rpart.plot: plot ‘rpart’ models: an enhanced version of ‘plot.rpart’; 2024. https://CRAN.R-project.org/package=rpart.plot [Accessed 4 May 2024].
32. Ripley B. Tree: classification and regression trees; 2023. https://CRAN.R-project.org/package=tree [Accessed 4 May 2024].
33. Breiman L, Curtler A. randomForest: Breiman and Curtler’s random forests for classification and regression; 2022. https://CRAN.R-project.org/package=randomForest [Accessed 4 May 2024].
34. Schloerke B, Cook D, Larmarange J, Briatte F, Marbach M, Thoen E, et al. GGally: extension to ‘ggplot2; 2024. https://CRAN.R-project.org/package=GGally [Accessed 4 May 2024].