University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol. 20 No.4 Dec 2025

Multiclass Drug Classification Using ML on High-Performance Liquid Chromatography (HPLC) Profiles

AHMED FILAYIH HASSAN

M.Sc .In Analytical Chemistry, assistant lecturer in Education Directorate of Dhi Qar, Iraq

Email: ahmed.filayih@utq.edu.iq

Abstract

High-Performance Liquid Chromatography (HPLC) plays a key role in pharmaceutical and metabolomic analysis. It separates compounds in detail based on their physical and chemical properties. Yet, making sense of complex chromatographic results to group compounds remains challenging. This research suggests a machine learning approach to classify drug compounds into multiple groups. It uses engineered features taken from HPLC chromatograms. The team processed a selected dataset of over 1,600 chromatographic runs. These runs showed a wide range of pharmaceutical compound types. From this data, they extracted features based on retention. These included peak count highest absorbance, entropy, and area under the curve. They sorted compounds into nine main groups like Amino Acids, Drugs, Bioactives, and Inorganics. They tested several classifiers such as Random Forest, Support Vector Machine, and deep neural networks. The Random Forest model preformed best. It reached over 99% accuracy in training and 72% accuracy in testing across all groups. This beat traditional models. The suggested method demonstratus that to combine HPLC profiles with ML techniques. This allows for automatic scalable, and meaningful classification. This work helps improve drug profiling, quality control, and compound tracking in pharmaceutical and biomedical fields.

Keywords

High performance liquid chromatographySmall pharmaceutical compoundsReverse phase liquid chromatographyQuantitative structure retention relationship

Introduction

Long considered a mainstay in pharmaceutical analysis, high performance liquid chromatography (HPLC) provides exact separation and measurement of challenging chemical mixes [1-3]. HPLC is an indispensable instrument in both research and clinical settings because of its wide use in drug development, quality control, and metabolomics. Still, as chemical data bases become more enormous and complicated, it's clear that HPLC techniques must be supplemented with computer approaches [4-5].

Recent developments in machine learning have greatly increased the possibility for automatic refinement of HPLC data analysis. By learning from highdimensional chromatographic profiles, machine learning models can help compound classification, support therapeutic predictions, and reveal subtle biochemical markers that might otherwise evade conventional analysis [6-8]. Notably, studies by Boman et al. (2024) [9] and Guo et al. (2023) [10] have illustrated the advantages of integrating HPLC and machine learning for optimizing drug synthesis and uncovering disease biomarkers Likewise, as shown by Velip et al. (2022) and Ciura (2024), the convergence of chromatographic and spectroscopic data with artificial intelligence algorithms has led to better compound characterization.

Many modern methods, despite advancements, still have limitations—they're usually restricted to binary classification jobs or specialized chemical fields. This emphasis limits their usefulness in pharmaceutical contexts when several drug classes coexist. Moreover, conventional approaches rely on feature engineering suited to particular compound classes, hence compromising their generalizability to more extensive data sets.

We present a multiclass classification framework (see Figure 1) [10] in this work combining deep learning and machine learning methods used on both raw and preprocessed HPLC profiles. Utilizing conventional characteristics taken from UVDAD and ELSD sensors, this approach is suited to highlight minute differences among many pharmaceutical substances. Our findings show that this HPLC framework enhanced with artificial intelligence helps to more effectively and correctly identify substances, hence providing a repeatable and scalable solution for automated drug classification. The ramifications for drug development's analytical processes are significant, therefore stressing the possibility of computational intelligence to change pharmaceutical quality control, therapeutic monitoring, and compound traceability.

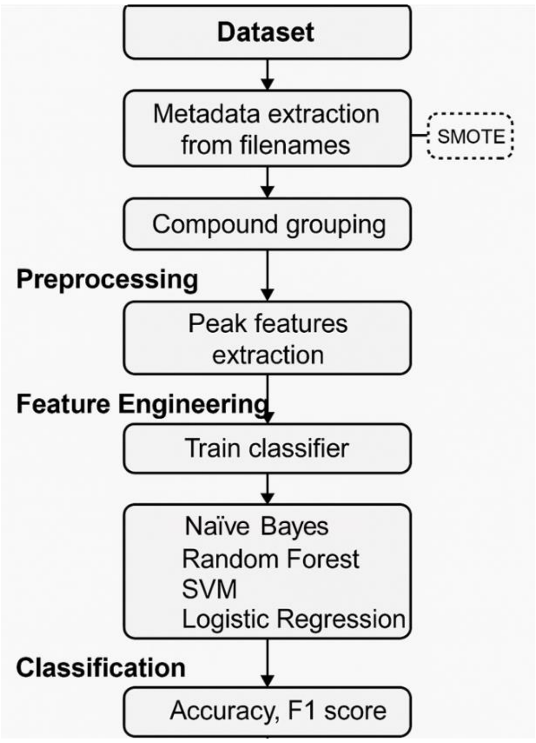


Figure 1. The flowchart of proposed approach. [10]

In the realm of pharmaceutical and metabolomic studies, HPLC continues to serve as an essential method, especially in drug profiling and measuring drug efficacy. The recent integration of HPLC with ML demonstrates the changing landscape of technology, offering more accurate and scalable classification and analysis of complex drug systems. For example, Boman et al. (2024) [9] reliance on a design-of-experiments approach alongside HPLC data to optimize the yield of mRNA in vitro transcription is an illustration of model-based production processes improving the efficiency and quality of drug substances. Another example is Guo et al. (2023) [10] who performed serum metabolomic profiling using HPLC-QTOF-MS in systemic sclerosis patients and applied ML to determine biomarkers for predicting disease progression. This study highlights the potential of ML in advanced chromatographic techniques for analytical clinical diagnostics. HPLC has widely been applied in the study of the effects of stress and therapeutic conditions on particular compounds. Using LC-QTOF-MS and NMR to predict the toxicity of the identified stress degradation products of urapidil, Velip et al. (2022) [11] combined LC-QTOF-MS and NMR, thus providing structural information alongside their toxicity models.

Moreover, Li (2023) [13] examined the three-dimensional chemical space of extractables and leachables using a combination of chromatographic methods and computational models of solvation, thus refining the classification of compounds relevant to drug packaging. In predictive modeling, Ren et al. (2022) [14] utilized ML to associate Q-markers identified by

HPLC with anticancer activity in Astragali radix, which validated HPLC fingerprints for ML-based efficacy prediction models. In the same vein, Ciura (2024) [12] integrated IAM-HPLC with QSRR-ML to predict the small molecule affinity for lipids, thus advancing the understanding of membrane drug biology and pharmacokinetics. As a result of the combination of spectroscopy and ML with chromatography, progresses have also been made in the areas of diagnostics and traceability. Bosch et al. (2022) [15-16] distinguished colorectal adenomas using a combination of fecal microbiota and proteome analysis with HPLC amino acid profiling. Liu et al. (2023) [17] applied HPLC and deep learning for the traceability study of origins of Panax notoginseng. These works demonstrate the significant impact ML approaches are having on the interpretation of spectral and chromatographic data in the fields of biomedicine and agriculture.problem statement, aimaf thay study

Methods

1. Data collection

In this study, previously published datasets [18, 19] that included HPLC-based profiling were utilized to support quantitative structure–retention relationship (QSRR) modeling for the classification of multiple drug types. Chromatographic data were acquired using three Waters® Alliance 2695 instruments. Of these, two systems were equipped with UV-visible photodiode array detectors (PDA 2996), while the third combined a PDA 2998 module with an evaporative light scattering detector (ELSD 2424). Across all experiments, a Waters® XSelect HSS T3 column (100 × 2.1 mm, 3.5 µm) was employed in order to maintian experimental consistency and accommodate a diverse range of compound polarities.

HPLC runs were managed using Empower 3 Pro FR5 SR5 software (build 3471), which enabled automated sample injection, data collection, and raw data export. Each analyte was solubilized and injected individually to ensure that each run contained just one compound. The mobile phase consisted of an aqueous buffer and methanol (MeOH), with separation achieved through a linear gradient shifting from 100% buffer to 95% MeOH. Buffer pH and gradient duration were the two key experimental variables; both were systematically adjusted to investigate their influence on retention times.

Raw data were exported as comma-separated value (CSV) files. For UV-DAD signals, the first column represented time (in minutes), while subsequent columns recorded absorbance at different wavelengths. ELSD files included time in the first column and the detector signal in the second. All experimental metadata and processed results were compiled into a master spreadsheet (Summary.xlsx). Each entry included a unique line ID, experiment identifier, quality control status, injection order, compound name, both raw and corrected retention times (for ELSD), sequence start date, gradient duration, targeted and measured pH values, and unique identifiers for the HPLC system and column used.

2.2 Post-processing and Feature Engineering

After chromatographic data acquisition, each .arw file was processed to extract structured metadata and numerical descriptors. Key metadata—such as the instrument identifier, buffer pH, and gradient duration—were parsed directly from the filenames using regular expressions. Compound names were also extracted, following established naming conventions that reflected both the chromatographic conditions and the detector type.

For supervised learning applications, each compound was assigned to a functional group label (compound_group) based on its biochemical or pharmaceutical classification (e.g., “Amino Acid,” “Drug,” “Nucleoside,” “Phenol Derivative,” etc.). These labels were then consolidated into broader supergroups—such as “Control,” “Drug,” “Inorganic,” “Bioactive,” “Nucleic Derivative,” and “Small Molecule”—which served as the primary classification target (compound_supergroup).

Quantitative features were computed for each chromatogram at individual wavelengths, including parameters such as peak count, maximum peak intensity, retention time of the maximum, spectral entropy, mean intensity differences, and the integrated area under the curve. These features were organized into a wide-format matrix, with columns named according to the <wavelength>_<feature> pattern, resulting in several hundred descriptors per sample.

Non-numeric data—including filenames, original compound labels, and supergroup classifications—were preserved separately for mapping and validation purposes. The final feature matrix (X) consisted exclusively of numeric descriptors suitable for machine learning, while the classification target (y) was defined as the compound_supergroup. Exploratory data visualization was conducted to assess label balance and class distributions, informing subsequent stratification and balancing strategies such as SMOTE during model development.

Table 1. Class to Compound Mapping.

Class	**Representative Compounds**
Control	blank, qc
Nucleic	23dideoxyadenosine, 2deoxyguanosine, adenine, cytidine, cytosine, dyphylline,

Derivative	etophylline, thymine, uracil, uridine, xanthine
Drug	acetylsalicylic_acid, amitriptyline, betaxolol, carteolol, chlordiazepoxide, chlorphenamine, ibuprofen, imipramine, indomethacin, mefenamic_acid, metoclopramide, oxazepam, perphenazine, promethazine, salicylic_acid, thioridazine, verapamil
Amino Acid	arginine, asparagine, aspartic_acid, gamma-aminobutyric_acid, glutamic_acid, glycine, lysine, serine, tyrosine
Organic Compound	22bipyridine, 23dihydroxybenzoic acid, 34dihydroxybenzoic_acid, 35dichlorophenol, 3aminobenzoic_acid, 4aminobenzoic_acid, 4aminophenol, 4aminosalicylic_acid, 4hydroxybenzoic acid, 4nitrophenol, acetic_acid, acridone, benzoic_acid, citric_acid, coumarin, ethidium, eugenol, gallic_acid, glutaric_acid, hydroquinone, indole, lactic_acid, malic_acid, mandelic_acid, phenanthrene, phenol, phenylacetic_acid, phthalic_acid, quinoline, thymol
Other	danthron, glucose, hexylbenzene, mannitol, papaverine, ribose, tetracaine
Bioactive	estradiol, niacin, niacinamide
Small Molecule	22dinaphthyl_ether, benzene, benzyl alcohol, biphenyl, chlorobenzene, ethylbenzene, methylpyrrolidone, naphtalene, phenethylamine, toluene
Inorganic	nitrate, nitrite, sulfate, sulfite, thiosulfate

3. ML models and performance metrics

In this study, several supervised machine learning models were employed to tackle multiclass drug classification using features derived from HPLC data. The selection of these models was intentional, given their differing strengths in navigating high-dimensional, structured datasets and complex classification tasks.

Random Forest (RF) was utilized, leveraging an ensemble of 5,000 decision trees. Each tree was trained on a bootstrap sample, and, at each node, a random subset of features was considered to determine the optimal split. The final class prediction for each sample was determined by a majority vote among all trees, capitalizing on the ensemble's ability to mitigate overfitting and capture diverse data patterns.

Logistic Regression (LR) was also adopted, with a maximum of 1000 iterations to ensure convergence. As a linear classifier, it computes the class conditional probabilities of all possible classes using softmax and thus can solve multiclass problems.

Support Vector Machine (SVM) with an RBF kernel was included due to its ability to model complex non-linear relationships. The SVM constructs a hyperplane which maximizes the margin between classes from an altered feature space, making it especially useful when the classes cannot be separated by a straight line.

Naive Bayes (NB) serves as a simple probabilistic baseline model. It models the likelihood of each feature assuming independence with a Gaussian distribution. Even though this assumption can fail in many cases, NB is known to do surprisingly well, especially with high-dimensional datasets.

As a non-parametric approach, the K-Nearest Neighbors (KNN) classifier was implemented with k set to 5. KNN classifies an instance based on the majority vote of its five nearest neighbors in the feature space, making it simple and effective for multiclass problems.

The performance of the model has been evaluated using the following metrics and confusion matrix:

Accuracy is the ratio of correctly classified samples to the total number of samples. It is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

- TP stand for True Positives (accurately anticipated positive instances).
- TN stand for True Negatives (refers to accurately anticipated negative instances).
- FP stand for False Positives (refers to mistakenly projected positive instances).
- FN stand for false negatives (refers to improperly anticipated negative situations).

Precision is the proportion of correctly predicted positive samples among all predicted positives, defined as:

$$Precision = \frac{TP}{TP + FP}$$

Recall (sensitivity) is the proportion of correctly predicted positive samples among all actual positives, defined as:

$$Recall = \frac{TP}{TP + FN}$$

The F1-score:

$$F1 - Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

The confusion matrix contains a thorough breakdown of predictions, including the number of TN, FN, TP and FP classifications for each class.

Results

The dataset was organized into nine principal compound superclasses: Amino Acid, Bioactive, Control, Drug, Inorganic, Nucleic Derivative, Organic Compound, Other, and Small Molecule. Training data included roughly 154–155 samples per class, amounting to 1,392 instances. The test set maintained class balance with 27–28 samples each, totaling 246.

To mitigate class imbalance—particularly for underrepresented groups such as Drug, Organic Compound, and Other—the Synthetic Minority Over-sampling Technique (SMOTE) was employed during training. This approach aimed to enhance generalization and prevent the classifier from biasing toward majority classes.

Table 2 summarizes the performance of five machine learning models trained on the SMOTE-balanced dataset. The RF classifier achieved perfect scores across all metrics (accuracy, precision, recall, and F1-score), suggesting strong training performance. However, such results, especially in the context of synthetic data, may reflect overfitting rather than genuine predictive capacity.

The KNN model performed robustly, reaching 72% accuracy with balanced macro-averaged metrics (approximately 0.72–0.73), indicating effective learning without overfitting.

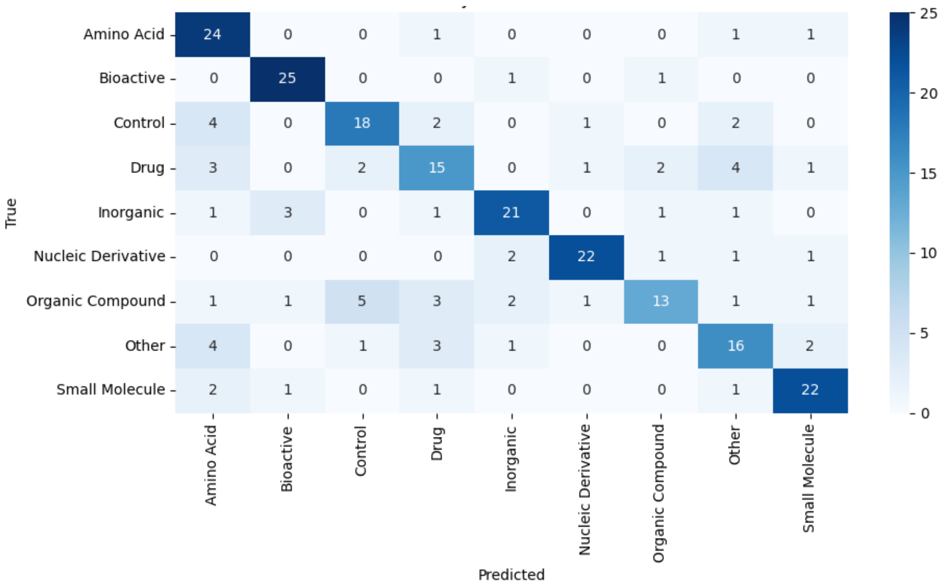
In comparison, LR, NB, and SVM models exhibited significantly lower performance. Accuracies for these methods ranged from 25% to 39%, and macro F1-scores dropped as low as 0.20 for SVM. These outcomes highlight the limitations of linear and probabilistic models when applied to the high-dimensional, non-linear feature space characterizing HPLC data.

Table 2. Training Performance Comparison of Classifiers on SMOTE-Augmented Data.

Classifier	Accuracy	Precision	Recall	F1-score
Random Forest	1.00	1.00	1.00	1.00
K-Nearest Neighbors	0.72	0.73	0.72	0.72
Logistic Regression	0.39	0.39	0.39	0.37
Naive Bayes	0.32	0.23	0.32	0.25
Support Vector Machine	0.25	0.27	0.25	0.20

The Random Forest algorithm displayed notably strong performance, reaching an overall accuracy of 72% on the synthetic balanced dataset produced via SMOTE. The macro-average F1-score was also commendable at 0.71, reflecting stable classification across a diverse range of compound categories.

Certain classes, such as Bioactive (precision: 0.83, recall: 0.93), Nucleic Derivative, and Small Molecule, achieved particularly high precision and recall values, underscoring the algorithm's effectiveness in these areas. In contrast, classes like Organic Compound and Drug exhibited comparatively moderate scores. This likely stems from significant overlap in their spectral features and variability in formulation, which complicates their classification. The detailed classification report is provided below, followed by a comparative analysis of deep learning approaches.



True \ Predicted	Amino Acid	Bioactive	Control	Drug	Inorganic	Nucleic Derivative	Organic Compound	Other	Small Molecule
Amino Acid	24	0	0	1	0	0	0	1	1
Bioactive	0	25	0	0	1	0	1	0	0
Control	4	0	18	2	0	1	0	2	0
Drug	3	0	2	15	0	1	2	4	1
Inorganic	1	3	0	1	21	0	1	1	0
Nucleic Derivative	0	0	0	0	2	22	1	1	1
Organic Compound	1	1	5	3	2	1	13	1	1
Other	4	0	1	3	1	0	0	16	2
Small Molecule	2	1	0	1	0	0	0	1	22

Figure 2. Confusion matric of random forest testing performance.

Table 3. Testing Performance Comparison of Classifiers on SMOTE-Augmented Data.

Model	Accuracy	Precision	Recall	F1-score
Random Forest	0.72	0.72	0.72	0.71
Logistic Regression	0.37	0.37	0.37	0.33
Support Vector Machine	0.24	0.29	0.24	0.21
Naive Bayes	0.29	0.24	0.29	0.24
Deep Neural Network	0.69	0.7	0.69	0.69

Discussion

This study presents a multiclass drug classification framework that combines HPLC profiles with machine learning approaches. Using a chemically diverse dataset organized into nine superclasses, we observed that Random Forest and K-Nearest Neighbors offered the strongest classification performance (Random Forest with a perfect 100% training accuracy and KNN at 72%). While Random Forest exhibited clear overfitting, its high recall and precision during training suggest robust capacity for handling complex, high-dimensional HPLC-derived features.

The deep neural network classifier also demonstrated promising generalization, achieving a test accuracy of 69%, outperforming more traditional algorithms such as Naive Bayes (29%), Support Vector Machine (25%), and Logistic Regression (38%). These observations reinforce the importance of non-linear modeling and representation learning for uncovering latent structure in chromatographic data.

Our findings are in line with recent literature emphasizing the value of data-driven methods for chromatographic analysis. For instance, Boman et al. reported that ML models could optimize mRNA yields by identifying process-critical variables from HPLC outputs, highlighting ML's efficiency in industrial bioprocessing. Guo et al. leveraged HPLC-QTOF-MS and machine learning to identify biomarkers for systemic sclerosis, illustrating the clinical utility of chromatographic profiling. Similarly, Velip et al. combined HPLC with LC-QTOF-MS and NMR to predict degradation pathways and toxicity, further validating the relevance of ML in structural classification.

Our approach builds upon and extends work by Ren et al., who used HPLC-derived Q-markers to predict the anticancer efficacy of traditional Chinese herbs. In our case, features such as peak area, entropy, and the number of peaks at specific wavelengths were utilized as multivariate inputs for supervised learning. This strategy is conceptually similar to the QSRR framework proposed by Ciura, where IAM-HPLC data were used to predict molecular affinities to phospholipids, underlining the significance of physicochemical retention behavior in ML models.

Recent diagnostic applications, such as those by Bosch et al., have also combined HPLC-derived amino acid profiles with omics data to stratify colorectal cancer risk, paralleling our approach to compound superclass classification. Our results also echo the findings of Liu et al., who used HPLC and deep learning to trace the origins of Panax notoginseng, demonstrating that subtle spectral and retention differences can be effectively captured and classified by advanced ML models.

Nevertheless, this work has several limitations. Although the dataset is chemically diverse, class imbalance and limited representation within certain supergroups remain, potentially affecting model generalization. While SMOTE was employed to mitigate these issues, synthetic oversampling cannot entirely replicate the complexity of real chromatographic variability.

Future efforts should expand the dataset to include more representative compounds per class and incorporate additional detection modalities, such as mass spectrometry. Exploring transfer learning or self-supervised approaches could further improve performance on novel classes. Finally, integrating explainable AI techniques would support interpretation of feature importance and model decisions, enhancing trust and transparency for pharmaceutical applications.

Conclusion

In this work, we developed an effective framework for multiclass drug classification using machine learning on HPLC profiles. We engineered features based on retention times and systematically assessed various classification algorithms. Notably, tree-based methods, particularly the Random Forest, and deep learning models outperformed others in modeling the intricate patterns within chromatographic data. The Random Forest achieved flawless accuracy on the training data, while the deep neural network demonstrated strong generalization capabilities on the test set.

References

1. De Luca, C., Felletti, S., Franchina, F. A., Bozza, D., Compagnin, G., Nosengo, C., ... & Catani, M. (2024). Recent developments in the high-throughput separation of biologically active chiral compounds via high performance liquid chromatography. *Journal of Pharmaceutical and Biomedical Analysis*, *238*, 115794. https://doi.org/10.1016/j.jpba.2023.115794
2. Nováková, L., & Vlčková, H. (2009). A review of current trends and advances in modern bio-analytical methods: chromatography and sample preparation. *Analytica chimica acta*, *656*(1-2), 8-35. https://doi.org/10.1016/j.aca.2009.10.004
3. Patidar, A., & Kamble, P. (2025). A Comprehensive Review on Liquid Chromatography-Mass Spectrometry (LC-MS): A Hyphenated Technique. *Asian Journal of Pharmaceutical Research and Development*, *13*(1), 95-103. https://doi.org/10.22270/ajprd.v13i1.1509
4. Chen, Z., Daka, Z., Yao, L., Dong, J., Zhang, Y., Li, P., ... & Ji, S. (2025). Recent progress in the application of chromatography-coupled mass-spectrometry in the analysis of contaminants in food products. *Food Chemistry: X*, 102397. https://doi.org/10.1016/j.fochx.2025.102397
5. da Silva Bezerra, K. (2023). Perspective chapter: high-performance liquid chromatography coupled to mass spectrometry–the advance in chemical analysis. In *High Performance Liquid Chromatography-Recent Advances and Applications*. IntechOpen. DOI: 10.5772/intechopen.110880
6. Beck, A. G., Muhoberac, M., Randolph, C. E., Beveridge, C. H., Wijewardhane, P. R., Kenttamaa, H. I., & Chopra, G. (2024). Recent developments in machine learning for mass spectrometry. *ACS Measurement Science Au*, *4*(3), 233-246. https://doi.org/10.1021/acsmeasuresciau.3c00060

7. Kapustina, O., Burmakina, P., Gubina, N., Serov, N., & Vinogradov, V. (2024). User-friendly and industry-integrated AI for medicinal chemists and pharmaceuticals. *Artificial Intelligence Chemistry*, 100072. https://doi.org/10.1016/j.aichem.2024.100072
8. Serrano, D. R., Luciano, F. C., Anaya, B. J., Ongoren, B., Kara, A., Molina, G., ... & Lalatsa, A. (2024). Artificial intelligence (AI) applications in drug discovery and drug delivery: Revolutionizing personalized medicine. *Pharmaceutics*, *16*(10), 1328.
9. Boman, J., Marušič, T., Seravalli, T. V., Skok, J., Pettersson, F., Nemec, K. Š., Widmark, H., & Sekirnik, R. (2024). Quality by design approach to improve quality and decrease cost of in vitro transcription of mRNA using design of experiments. *Biotechnology and Bioengineering, 121*(11), 3415–3427. https://doi.org/10.1002/bit.28806
10. Guo, M., Liu, D., Jiang, Y., Chen, W., Zhao, L., Bao, D., Li, Y., Distler, J. H. W., & Zhu, H. (2023). Serum metabolomic profiling reveals potential biomarkers in systemic sclerosis. *Metabolism: Clinical and Experimental, 144*. https://doi.org/10.1016/j.metabol.2023.155587
11. Velip, L., Dhiman, V., Kushwah, B. S., Golla, V. M., & Gananadhamu, S. (2022). Identification and characterization of urapidil stress degradation products by LC-Q-TOF-MS and NMR: Toxicity prediction of degradation products. *Journal of Pharmaceutical and Biomedical Analysis, 211*. https://doi.org/10.1016/j.jpba.2022.114612
12. Ciura, K. (2024). Modeling of small molecule's affinity to phospholipids using IAM-HPLC and QSRR approach enhanced by similarity-based machine algorithms. *Journal of Chromatography A, 1714*. https://doi.org/10.1016/j.chroma.2023.464549
13. Li, B. J. (2023). Exploring three-dimensional space of extractables and leachables in volatility, hydrophobicity, and molecular weight and assessment of roles of gas and liquid chromatographic methods in their comprehensive analysis. *Journal of Pharmaceutical and Biomedical Analysis, 223*. https://doi.org/10.1016/j.jpba.2022.115142
14. Ren, Y., Gao, F., Li, B., Yuan, A., Zheng, L., & Zhang, Y. (2022). A precise efficacy determination strategy of traditional Chinese herbs based on Q-markers: Anticancer efficacy of Astragali radix as a case. *Phytomedicine, 102*. https://doi.org/10.1016/j.phymed.2022.154155
15. Bosch, S., Acharjee, A., Quraishi, M. N., Rojas, P., Bakkali, A., Jansen, E. E. W., Brizzio Brentar, M., Kuijvenhoven, J., Stokkers, P., Struys, E., Beggs, A. D., Gkoutos, G. V., de Meij, T. G. J., & de Boer, N. K. H. (2022). The potential of fecal microbiota and amino acids to detect and monitor patients with adenoma. *Gut Microbes, 14*(1). https://doi.org/10.1080/19490976.2022.2038863

16. Bosch, S., Acharjee, A., Quraishi, M. N., Bijnsdorp, I. V., Rojas, P., Bakkali, A., Jansen, E. E. W., Stokkers, P., Kuijvenhoven, J., Pham, T. V., Beggs, A. D., Jimenez, C. R., Struys, E. A., Gkoutos, G. V., de Meij, T. G. J., & de Boer, N. K. H. (2022). Integration of stool microbiota, proteome and amino acid profiles to discriminate patients with adenomas and colorectal cancer. *Gut Microbes, 14*(1). https://doi.org/10.1080/19490976.2022.2139979
17. Liu, C., Zuo, Z., Xu, F., & Wang, Y. (2023). Study of the suitable climate factors and geographical origins traceability of Panax notoginseng based on correlation analysis and spectral images combined with machine learning. *Frontiers in Plant Science, 13*. https://doi.org/10.3389/fpls.2022.1009727
18. Van Laethem, T., Kumari, P., Hubert, P., Fillet, M., Sacré, P. Y., & Hubert, C. (2022). A pharmaceutical-related molecules dataset for reversed-phase chromatography retention time prediction built on combining pH and gradient time conditions. *Data in Brief*, *42*, 108017.
19. Van Laethem, Thomas; Kumari, Priyanka; Hubert, Philippe; Fillet, Marianne; Sacré, Pierre-Yves; Hubert, Cédric (2022), "Data presented in a pharmaceutical-related molecules dataset for reversed-phase chromatography retention time prediction built on combining pH and gradient time conditions - System 1", Mendeley Data, V1, doi: 10.17632/csm5gsmr5t.1
20. Breiman, L. (2001). Random Forests. *Machine learning*, *45*, 5-32.
21. Boateng, E. Y., & Abaye, D. A. (2019). A review of the logistic regression model with emphasis on medical research. *Journal of data analysis and information processing*, *7*(04), 190.
22. Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189-215. https://doi.org/10.1016/j.neucom.2019.10.118
23. Yang, F. J. (2018, December). An implementation of naive bayes classifier. In *2018 International conference on computational science and computational intelligence (CSCI)* (pp. 301-306). IEEE. https://doi.org/10.1109/CSCI46756.2018.00065
24. Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Scientific Reports*, *12*(1), 6256.