IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

MACHINE LEARNING FOR BIOMARKER DISCOVERY IN METABOLOMICS: TECHNIQUES, APPLICATIONS, AND FUTURE DIRECTIONS

[1]Siddhant Yadav, [2]Abhishek Kumar Gautam

[1]Department of Biotechnology, Amity University, Lucknow, Uttar Pradesh, India

[2] School of Studies in Forensic Science, Vikram University, Ujjain, Madhya Pradesh, India

ABSTRACT

The application of machine learning (ML) and artificial intelligence (AI) in metabolomics provides revolutionary possibilities for improving disease diagnostics and comprehending intricate biochemical processes. The study of small molecules in biological systems, known as metabolomics, gains from AI/ML's capacity to analyse enormous datasets and identify complex patterns and biomarkers. The main developments in AI-driven metabolomics are highlighted in this review, with particular emphasis on the use of ML models such as random forests, support vector machines, and artificial neural networks for disease diagnosis and biomarker discovery, as well as deep learning approaches for metabolite identification and multi-omics data integration. Even with this tremendous advancement, issues like population variability management, repeatability, and interpretability of models still need to be resolved. It is imperative that these problems are resolved in order to advance personalised medicine and realise dependable clinical metabolomics applications.

Keywords: Machine Learning, Metabolomics, Artificial Intelligence, Biomarker, Omics, Medicine

I. INTRODUCTION

1. Metabolomics

The study of small metabolites or chemical reactions involving tiny substrates in tissues or organisms is known as metabolomics. All the metabolites found in any biological cells, tissue, or organ, as well as the cellular products that result from them, are represented by the metabolome. It can be used to investigate biological data at the biochemical level and gives a snapshot of the physiology of the cell under study. This offers a line of inquiry into the biological phenotype that can be utilized to comprehend health and illness [1]In the late 1940s, Roger Williams developed the idea of a metabolic profile [2]. Using paper chromatography, he proposed that distinct metabolic patterns in saliva and urine are indicative of schizophrenia. The term "metabolic profile" was not coined until the 1970s, when mass spectrometry and gas chromatography were introduced, along with other technological breakthroughs [3]. The Metabolite and Chemical Entity Database (METLIN), the first comprehensive database for metabolomic tandem mass spectrometry, was created in 2005 at the Scripps Research Institute [4] Under the direction of David S. Wishart, "The Human Metabolome Project" created the initial draft of a database in 2007 that contained around 2,500 metabolites, 1,200 medications, and approximately 3,500 food components. The ability of methods like mass spectrometry and gas chromatography to identify thousands of distinct characteristics in a single specimen has advanced, making the computational problem of finding metabolites linked to a disease or trait more and more challenging. A thorough evaluation of biological specimens and the molecules they are connected with has been made possible by the field of metabolomics. According to Gowda et al. (2008), a better understanding of the biological system at the molecular level is essential for disease diagnosis and the development of new treatments. Metabolomics is the fundamental layer in the area of omics that captures all the information expressed and changed by the genetic regulatory and processing layers upstream. The closest connection to the phenotypic is this. Because of its direct relevance to the field of biomarker discovery, it is at the forefront of personalised health in terms of diagnosis and treatment. Because biological systems are intricate, understanding them frequently necessitates integrating multiple layers of omic data. As the result of the interaction between the different omic layers, metabolomics is a possible remedy for this[5].

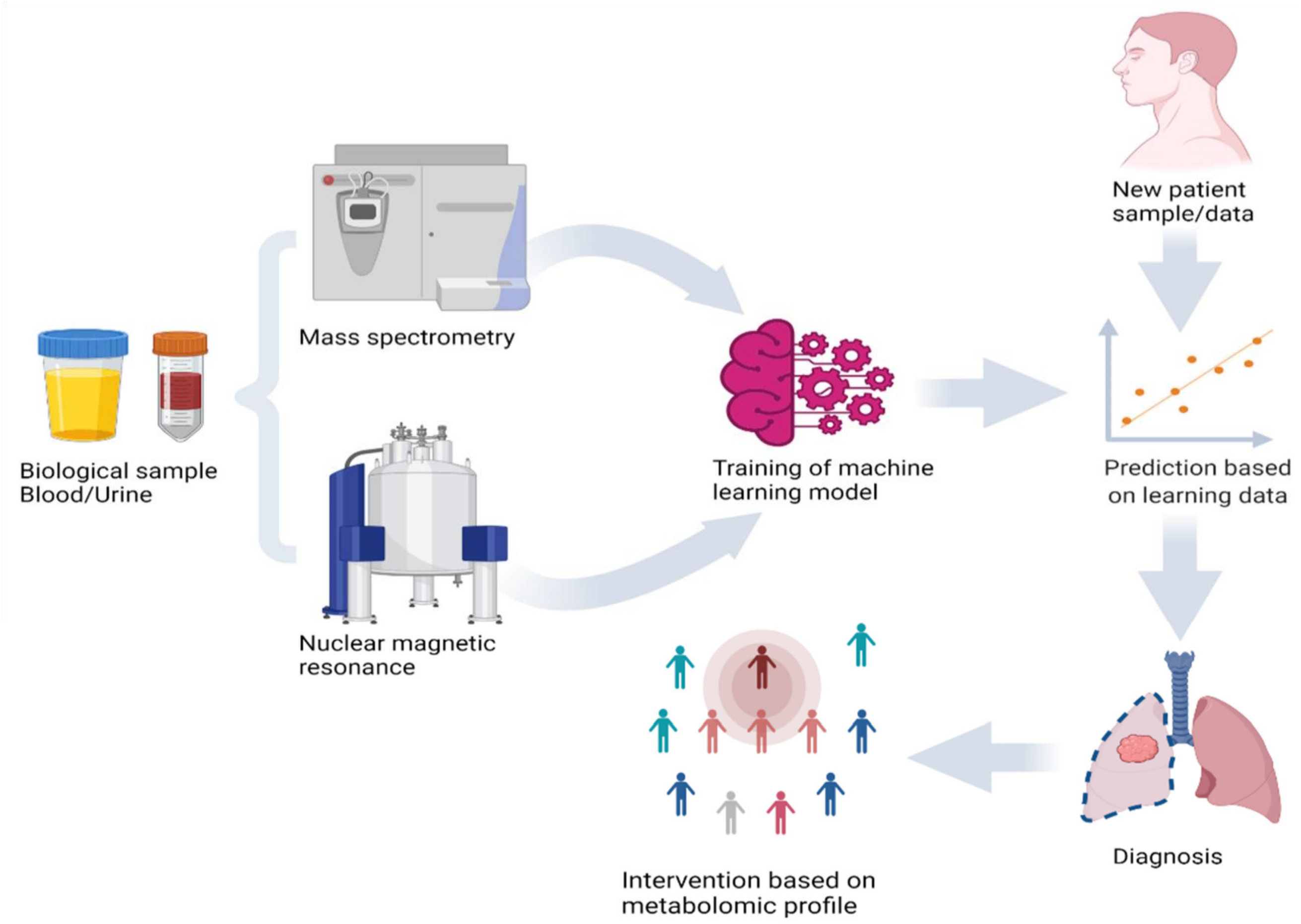
Biological sample
Blood/Urine
Mass spectrometry
Nuclear magnetic resonance
Training of machine learning model
New patient sample/data
Prediction based on learning data
Diagnosis
Intervention based on metabolomic profile

Figure 1: A Machine learning model showing prediction of samples using metabolomics analysis of biological samples with mass spectrometry and nuclear magnetic resonance.

2. Machine Learning

The concept and area of artificial intelligence (AI) have garnered significant attention in the 21st century. AI and machine learning (ML) offer endless possibilities with their many applications in comprehending the structures or trends in enormous amounts of data generated from contemporary high-throughput research. ML is used to create models that can process massive amounts of data and, via learning, resolve challenging issues. In the twenty-first century, artificial intelligence (AI) as a concept and field have attracted a lot of interest. With its numerous applications in understanding the structures or trends in massive amounts of data generated from modern high-throughput research, artificial intelligence (AI) and machine learning (ML) present countless opportunities. ML is used to build models that can handle large volumes of data and, through learning, solve difficult problems. A dataset used to create a machine learning model is typically split into two subsets: a testing subset, which contains about 30% of the data and is used to provide an unbiased assessment of the final model from the training step, and a training subset, which contains about 70% of the available data and is used in the ML algorithm to build a model and make predictions. For the machine learning algorithm to have more chances to learn and refine the model, a large amount of data must be used in the initial learning process. Formally, the algorithm is capable of learning through a mathematical function that associates particular inputs with particular outputs. Without being specifically designed, the algorithms utilise the training dataset as a reference to make predictions. This is accomplished by a sequence of processes in which learning is done using weights and biases to provide predictions in a finite number of steps [6].

There are three types of machine learning: semisupervised learning, unsupervised learning, and supervised learning. Using highly statistical techniques, supervised machine learning algorithms train a model on labelled data and generate predictions about unknown (unlabeled) data. Unsupervised models, on the other hand, use unlabeled training data. Because machine learning can handle both linear and non-linear data, it is the best application for mass spectrometry data. Yet, applying machine learning to mass spectrometry applications is not a new idea; a 1990s studyThough not a new idea, applying machine learning to mass spectrometry applications was first shown in a 1990s study that showed how well artificial neural networks (ANNs) could categorise mass spectra [7]. Subsequently, a multitude of additional supervised algorithms were employed to enhance the mass spectrometry data categorization process [8]. Machine learning-based diagnostic research in mass spectrometry started to expand in the twenty-first century.

II. APPLICATIONS

1. Machine Learning Applications for Biomarker Discovery

The correlation of variables makes the process of discovering biomarkers through machine learning complex and full of challenges. Several research have demonstrated how feature selection algorithms can be used to use metabolomics data to identify disease biomarkers. For instance, to find the most discriminating characteristics to use as potential biomarkers, researchers employed random forests with feature significance functions [9, 10]. They employed the top-ranking features to train the classifier model after calculating the relevance score assigned to each feature based on the Gini index calculation.

In a different study, the prospect of employing targeted metabolomics to analyse plasma samples to find biomarkers for lung cancer disease was explored. Quick correlation-based selection algorithms found five best-performing biomarkers that might distinguish lung cancer patients from healthy individuals [11]. In order to predict the state of renal cell carcinoma, Bifarin et al. [12] analysed urine samples using liquid chromatography–mass spectrometry and NMR. They also created a biomarker panel of 10 compounds by using the PLS regression approach and recursive feature selection. Following the selection of features based on their frequency of appearance in both feature selection techniques, 10 metabolites were chosen to train the classification model.

In a different investigation, the authors assessed serum samples examined by high-resolution mass spectrometry-based metabolomics to see how well multivariate approaches with unbiased variable selection (MUVR) performed. Using random forest, SVM, and logistic regression techniques, the MUVR approach identified 13 metabolites that yielded good results and created a panel of potential biomarkers that can differentiate gout from asymptomatic hyperuricemia [13]. Furthermore, PLS-DA has been employed to determine the most discriminant lipids and metabolites to distinguish serum metabolomic and lipidomic profiles of patients with rheumatoid arthritis from healthy controls [9]. Three classifier methods were used to assess the selected properties of the 26 molecules that the authors proposed as potential biomarkers for rheumatoid arthritis: logistic regression, random forest, and SVM.

2. Application of Machine Learning for the Diagnosis of Diseases

The number of metabolomics research that have used machine learning techniques has significantly increased since the year 2000. Numerous research have demonstrated the ability of machine learning to distinguish between groups that are healthy and those that are sick, as well as to find significant biomarkers that may be used in a range of clinical decision-making contexts [14, 15]. The most current uses of supervised machine learning for illness diagnosis are shown in the sections that follow.

2.1 Random Forest

Random forest is one of the most widely used supervised machine learning algorithms for mass spectrometry data due to its ability to cope with missing values, data noise, and reduced overfitting risk [15]. The ensemble methodology of decision trees is incorporated into the Random Forest classification and regression technique to predict classes. The decision trees' predictions are used by the algorithm to determine the result. The class that the majority of the trees choose for classification tasks is the random forest's output; each tree may be thought of as an uncorrelated model.

Compared to other classifier models, random forest was demonstrated to be more effective at selecting potential biomarkers, stability, prediction ability, and overfitting, and in diagnosing colorectal cancer (with 100% accuracy) using metabolomics data [16]. By using metabolomics data to identify the Zika virus, Melo et al. [9] demonstrated the robustness and better performance of random forest compared to other classifier models, proving that random forest works better than other classifier models. To distinguish across groups, 42 features were used in the model's development and evaluation. Lima et al. [18] used a mix of metabolomics and random forest data to claim 97% accuracy in diagnosing Para coccidioidomycosis.

Using random forest and metabolomics data, recent research has shown that malignant mesothelioma may be identified with 92% accuracy in the validation dataset [19]. Researchers investigated twenty dysregulated characteristics that set the malignant mesothelioma group apart from others. Biliverdin and bilirubin were shown to have diagnostic potential among the 20 characteristics. To further illustrate how random forest can be used to choose the best possible biomarkers, biliverdin was evaluated as the fourth-most significant variable overall by random forest. However, to obtain a precise picture of the diagnostic model, study constraints were also disclosed, such as a reduced number of classes and sample sizes.

Fukui et al. [20] combined logistic regression and random forest to achieve a higher score in terms of sensitivity and specificity than if each method had been employed alone in a study that focused on identifying irritable bowel syndrome. Using metabolomics data, four classifier models were developed by other researchers [21]: a generalised linear regression model, PLS-DA, PCA linear discriminant analysis, and random forest. Two methods were used to train the models. Using every metabolite, the models were trained in the first method. In the second method, only pre-selected metabolites were used to train the models. With an AUC score of 72%, the random forest model with pre-selected variables proved to be the most successful.

2.2 Support Vector Machine

Currently, the most popular machine learning method in precision medicine is SVM classification. SVM is a model that builds a decision boundary (hyper-plane) in a high-dimensional feature space using "support vectors." Datapoints that are near to the hyperplane are known as support vectors, and they help to optimise the hyperplane itself [22]. With as few data points as possible on the wrong side of the decision boundary, the hyperplane's goal is to maximise the distance between two classes [23, 24].

In order to maximise the distance for a given set of training samples, a hyperplane is created, which is expressed mathematically as

$$WTX + b = 0,$$

where W stands for weight matrix, X for dataset, and b for constant term.

SVM can also be used, via a technique known as the kernel trick, to classify non-linear data. The polynomial kernel, Gaussian kernel, Gaussian radial basis function (RBF), Laplace RBF, sigmoid kernel, hyperbolic tangent kernel, and linear splines kernel in one dimension are a few examples of the various kinds of kernel tricks utilised for various situations. Radial basis function (RBF), however, is the preferred option among other kernels and is frequently utilised for non-linear tasks in metabolomics. SVM was applied in a recent study to distinguish gout from asymptomatic hyperuricemia. The technique was utilised as a classifier in conjunction with random forest and logistic regression. According to the author, random forest performed better in the training set but worse in the test set, suggesting that the classifier model was overfitting. In contrast, SVM beat the other classifiers in terms of obtaining a higher area under curve score in the validation set.

Using a mouse model, Song et al. [25] employed SVM to find an early indicator of diabetes cognitive deficits. Using seven features, SVM was able to identify two sets of samples with 91.66% accuracy. The scientists also suggested biomarkers that may be involved in pathogenesis, such as the metabolism of nicotinamide and glutathione, tryptophan, and sphingolipids.

SVM was employed in a different recent work [26] to categorise Staphylococcus aureus multidrug resistance and benzylpenicillin resistance. The researchers used matrix-assisted laser desorption/ionization–time of flight mass spectrometry to find antibiotic resistance signature profiles in isolates of S. aureus. SVM outperformed naive Bayes, random forest, and multilayer neuron perceptron neural networks in terms of accuracy, specificity, and sensitivity.

Major depression was diagnosed by Zheng et al. [27] using the least-squares SVM (LS-SVM) with three kernel functions: linear, polynomial, and radial basis. With a radial basis function, LSSVM performed better on the test dataset than other kernel functions, achieving 96% accuracy. Glucose–lipid signalling characteristics, including polyunsaturated fatty acids, lipids containing acetoacetate, lipids with N-acetyl, glucose, adipic acid, and sugars including amino acids, were incorporated into the classifier's construction. SVM is a very appealing algorithm to conduct precision medicine studies and to find possible metabolic biomarkers, as all of the published research has demonstrated. Furthermore, SVM is especially useful when there are few biological replicates or patients.

2.3 Artificial Neural Networks

ANNs are capable of handling complex (non-linear) aspects inside input and making future situation predictions, just like the human brain. While ANNs learn by modifying the connections between the processing units that comprise the network structure, humans learn by making small changes to synaptic linked neurons.

Artificial neurons were defined in 1943 by McCulloch and Pitts [28] as a mathematical function created by imitating the functions of real biological neurons. The complexity of the scenario determines how many hidden layers and neurons there are in each layer. The external system sends a vector of predictor variables, each represented by a node, to the input layer. The first hidden layer's weights are then multiplied by these data, making them changed. These products are combined and sent through a non-linear transfer function (sigmoid, hyperbolic tangent) to provide an output that resembles an axon. In ANN-supervised learning, the weights are adjusted to roughly reflect each goal as a nonlinear function of the inputs.

An output layer with a back propagation technique, seven hidden layer units, and thirteen input layers were utilised by ANNs with an output layer in a metabolomics study conducted on plasma samples from Parkinson's disease patients [29]. The neural network algorithm's accuracy in identifying the course of the disease was 97.14%. One incidence of misdiagnosis was also noted, though.

Another study used flow infusion electrospray ion mass spectrometry to profile the compounds in the sputum of lung cancer patients and age-matched smokers as a control [30]. ANNs were then used to analyse the metabolomic profiles and diagnose lung cancer, namely non-small-cell lung cancer (NSCLC) and small-cell lung cancer (SCLC). The authors achieved an 80% sensitivity and 100% specificity in their classification of SCLC. Between the NSCLC and SCLC groups, six metabolites—phenylacetic acid, L-fucose, caprylic acid, acetic acid, propionic acid, and glycine—were found to represent potential indicators.

Additionally, ANNs were employed to elaborate metabolite abundances from dried blood spots that were subjected to direct infusion mass spectrometry analysis [31]. Acute cerebral infarction and intracerebral haemorrhage were contrasted by the writers. They were able to distinguish between intracerebral haemorrhage and acute cerebral infarction with over 70% accuracy when using an external validation set. They employed ANNs with 11 units and 10 hidden layers of neurons, and after training the ANN model with stepwise regression, they were able to identify 11 important metabolites.

III. Machine Learning Tools for Metabolomic Analysis

With its rapid advancement, machine learning (ML) now provides a broad range of methods to address complex problems in the field of metabolomics and identify putative biomarkers. Machine learning has already been applied to enhance data processing techniques in the NMR and mass spectrometry domains [32]. Numerous machine learning tools, including WEKA, KNIME, and Orange include user-friendly interfaces, don't require programming knowledge, and are available as open-source software [33]. Other well-liked open-source libraries are TPOT, which offers an automated pipeline of machine learning algorithms that uses genetic programming stochastic global search approach to sort out top-performing ML models [35], and Scikit-learn (also known as sklearn), which is used to implement machine learning algorithms in Python [34]. Numerous machine learning algorithms and feature selection techniques are available in the R Caret library [36]. Furthermore, the field of metabolomics is increasingly using ANN-based data processing tools such as PyTorch, Keras, and TensorFlow [35]. Yet, metabolomics data have not yet been analysed using recently established automated machine learning and deep learning pipelines as AutoGluon, AutoPrognosis, H2O, and PennAI [37,38]. The most popular tools and libraries for metabolomics research are listed in Table 1.

Table 1. Tools used by metabolomics studies for machine learning algorithms.

Tools/Libraries	Purpose of Use in Studies	Programing Language	Programing Skills Requirement	Metabolomic Studies
Weka	Classification/feature selection	Java	No	[33,37]
KNIME	Data processing	Java	No	[39,40]
Orange data mining	Classification	Python, Cython, C++, C	No	[41]
Scikit-learn	Data processing/Classification	Python	Yes	[12,42]
TPOT	Classification/feature selection	Python	Yes	[43]
Caret	Classification/feature selection	R	Yes	[44]
Keras and Tensor flow	Data processing/Peak identification	Python, R	Yes	[45]

IV. Limitations

Even with machine learning's progress in the healthcare industry, a number of issues still need to be addressed. The first challenge in creating a metabolomics-based machine learning-assisted diagnostic model is figuring out how little information is required to accurately depict a given biological issue or disease. Generating datasets with enough samples for training and assessing a strong model on a separate dataset, while also accurately representing the variance of the population, is a challenging task. In addition, biases in the experimental process may affect the quality of the metabolomics data. The reproducibility of outcomes and the algorithms' ability to explain them are two further issues with machine learning models; with larger datasets, it gets harder to understand the reasoning behind algorithmic choices [46]. Because the mathematics of prediction is based on incomplete knowledge, many algorithms are referred to as "black boxes". When it comes to applying machine learning to clinical decision making, this feature is a significant drawback.

Numerous classifiers have been employed for the identification of biomarkers as well as the diagnosis of illnesses, as demonstrated in earlier sections. Random forest demands a lot of processing power for large datasets, but it is typically less prone to overfitting. SVM, on the other hand, works incredibly well with high-dimensional data, although overfitting is a common side consequence. Lastly, while ANNs perform poorly with small datasets, they are especially well suited for large datasets.

V. CONCLUSION

In metabolomics and other high-throughput technologies, artificial intelligence is a commonly employed method, particularly for early detection. The effectiveness of machine learning in the field of medical science has been demonstrated by recent studies that have produced multiple algorithms for disease classification based on metabolomics profiles. However, there are still several challenges to be solved, such as how to interpret machine learning models and create reliable models that take population variability and disease into consideration. There is currently no gold standard for choosing the best algorithm to apply given a certain dataset. In fact, depending on the technique used, even little modifications to the dataset format can produce wildly different results.

VI. REFERENCES

[1] Gowd, G. A. N., Zhang, S., Gu, H., Vincent, A., Shanaiah, N., and Raftery, D.(2008). Metabolomics-based methods for early disease diagnostics. Expert Rev. Mol. diagn. 8 (5), 617–633. doi:10.1586/14737159.8.5.617

[2] Gates, S. C., and Sweeley, C. C. (1978). Quantitative metabolic profiling based on gas chromatography. Clin. Chem. 24 (10), 1663–1673. doi:10.1093/clinchem/24.10.1663

[3] Griffiths, W. J., and Wang., Y. (2009). Mass spectrometry: From proteomics to metabolomics and lipidomics. Chem. Soc. Rev. 38(7), 1882–1896. doi:10.1039/b618553n

[4] Smith, C. A., Grace O'Maille, E. J. W., Qin, C., Trauger, S. A., Brandon, T. R., Custodio, D. E., et al. (2005). Metlin: A metabolite mass spectral database. Ther. Drug Monit. 27 (6), 747–751. doi:10.1097/01.ftd.0000179845.53213.39

[5] Hasin, Y., Marcus, S., and Lusis, A. (2017). Multi-omics approaches to disease.Genome Biol. 18 (1), 83. doi:10.1186/s13059-017-1215-1

[6] Cohen, S. (2021). "Chapter 1 - the evolution of machine learning: Past, present, and future," in Artificial intelligence and deep learning in pathology. 1–12. Editor S. Cohen (Elsevier).

[7] Curry, B.; Rumelhart, D.E. Msnet: A Neural Network which Classifies Mass Spectra. Tetrahedron Comput. Methodol. 1990,

3, 213–237. [CrossRef]

[8] Broadhurst, D.I.; Goodacre, R.; Jones, A.; Rowland, J.J.; Kell, D.B. Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry. Anal. Chim. Acta 1997,348, 71–86. [CrossRef]

[9] Melo, C.F.O.R.; Navarro, L.C.; de Oliveira, D.N.; Guerreiro, T.M.; Lima, E.d.O.; Delafiori, J.; Dabaja, M.Z.; Ribeiro, M.d.S.;

de Menezes, M.; Rodrigues, R.G.M.; et al. A Machine Learning Application Based in Random Forest for Integrating Mass

Spectrometry-Based Metabolomic Data: A Simple Screening Method for Patients With Zika Virus. Front. Bioeng. Biotechnol. 2018,6, 31. [CrossRef] [PubMed]

[10] Dias-Audibert, F.L.; Navarro, L.C.; de Oliveira, D.N.; Delafiori, J.; Melo, C.F.O.R.; Guerreiro, T.M.; Rosa, F.T.; Petenuci, D.L.;Watanabe, M.A.E.; Velloso, L.A.; et al. Combining Machine Learning and Metabolomics to Identify Weight Gain Biomarkers. Front. Bioeng. Biotechnol. 2020, 8, 6. [CrossRef] [PubMed]

[11]. Xie, Y.; Meng,W.-Y.; Li, R.-Z.;Wang, Y.-W.; Qian, X.; Chan, C.; Yu, Z.-F.; Fan, X.-X.; Pan, H.-D.; Xie, C.; et al. Early lung cancer diagnostic biomarker discovery by machine learning methods. Transl. Oncol. 2021, 14, 100907. [CrossRef] [PubMed]

[12]. Bifarin, O.O.; Gaul, D.A.; Sah, S.; Arnold, R.S.; Ogan, K.;Master, V.A.; Roberts, D.L.; Bergquist, S.H.; Petros, J.A.; Fernández, F.M.; et al. Machine Learning-Enabled Renal Cell Carcinoma Status Prediction UsingMultiplatformUrine-Based Metabolomics. J. Proteome Res. 2021, 20, 3629–3641. [CrossRef]

[13]. Shen, X.; Wang, C.; Liang, N.; Liu, Z.; Li, X.; Zhu, Z.-J.; Merriman, T.R.; Dalbeth, N.; Terkeltaub, R.; Li, C.; et al. Serum

Metabolomics Identifies Dysregulated Pathways and Potential Metabolic Biomarkers for Hyperuricemia and Gout. Arthritis Rheumatol. 2021, 73, 1738–1748. [CrossRef]

[14]. Haq, A.U.; Li, J.P.; Memon, M.H.; Nazir, S.; Sun, R. A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms. Mob. Inf. Syst. 2018, 2018, 3860146. [CrossRef]

[15]. Mishra, V.; Singh, Y.; Rath, S.K. Breast Cancer detection from Thermograms Using Feature Extraction and Machine Learning Techniques. In Proceedings of the 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, 29–31 March 2019.

[16]. Amaratunga, D.; Cabrera, J.; Lee, Y.-S. Enriched random forests. Bioinformatics 2008, 24, 2010–2014. [CrossRef]

[17]. Chen, T.; Cao, Y.; Zhang, Y.; Liu, J.; Bao, Y.;Wang, C.; Jia,W.; Zhao, A. Random Forest in Clinical Metabolomics for Phenotypic Discrimination and Biomarker Selection. Evid.-Based Complementary Altern. Med. 2013, 2013, 298183. [CrossRef]

[18] Lima, E.d.O.; Navarro, L.C.; Morishita, K.N.; Kamikawa, C.M.; Rodrigues, R.G.M.; Dabaja, M.Z.; de Oliveira, D.N.; Delafiori, J.;Dias-Audibert, F.L.; Ribeiro, M.d.S.; et al. Metabolomics and Machine Learning Approaches Combined in Pursuit for More Accurate Paracoccidioidomycosis Diagnoses. mSystems 2020, 5, e00258-20. [CrossRef] [PubMed]

[19]. Li, N.; Yang, C.; Zhou, S.; Song, S.; Jin, Y.; Wang, D.; Liu, J.; Gao, Y.; Yang, H.; Mao, W.; et al. Combination of Plasma-Based Metabolomics and Machine Learning Algorithm Provides a Novel Diagnostic Strategy for Malignant Mesothelioma. Diagnostics 2021, 11, 1281. [CrossRef]

[20]. Fukui, H.; Nishida, A.; Matsuda, S.; Kira, F.; Watanabe, S.; Kuriyama, M.; Kawakami, K.; Aikawa, Y.; Oda, N.; Arai, K.; et al. Usefulness of Machine Learning-Based Gut Microbiome Analysis for Identifying Patients with Irritable Bowels Syndrome. J. Clin. Med. 2020, 9, 2403. [CrossRef] [PubMed]

[21]. Kasakin, M.F.; Rogachev, A.D.; Predtechenskaya, E.V.; Zaigraev, V.J.; Koval, V.V.; Pokrovsky, A.G. Targeted metabolomics

approach for identification of relapsing–remitting multiple sclerosis markers and evaluation of diagnostic models. MedChemComm 2019, 10, 1803–1809. [CrossRef]

[22]. Scölkopf, B.; Smola, A.J. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond; The MIT Press: Cambridge, MA, USA, 2018.

[23]. Huang, X.; Xu, Q.-S.; Yun, Y.-H.; Huang, J.-H.; Liang, Y.-Z. Weighted variable kernel support vector machine classifier for

metabolomics data analysis. Chemom. Intell. Lab. Syst. 2015, 146, 365–370. [CrossRef]

[24]. Mendez, K.M.; Reinke, S.N.; Broadhurst, D.I. A comparative evaluation of the generalised predictive ability of eightmachine learning algorithms across ten clinicalmetabolomics data sets for binary classification.Metabolomics 2019, 15, 150. [CrossRef] [PubMed]

[25]. Song, L.; Zhuang, P.; Lin, M.; Kang, M.; Liu, H.; Zhang, Y.; Yang, Z.; Chen, Y.; Zhang, Y. Urine Metabonomics Reveals Early Biomarkers in Diabetic Cognitive Dysfunction. J. Proteome Res. 2017, 16, 3180–3189. [CrossRef] [PubMed]

[26]. Esener, N.; Maciel-Guerra, A.; Giebel, K.; Lea, D.; Green, M.J.; Bradley, A.J.; Dottorini, T. Mass spectrometry and machine learning for the accurate diagnosis of benzylpenicillin and multidrug resistance of Staphylococcus aureus in bovine mastitis. PLoS Comput. Biol. 2021, 17, e1009108. [CrossRef] [PubMed]

[27].. Zheng, H.; Zheng, P.; Zhao, L.; Jia, J.; Tang, S.; Xu, P.; Xie, P.; Gao, H. Predictive diagnosis of major depression using NMR-based metabolomics and least-squares support vector machine. Clin. Chim. Acta 2017, 464, 223–227. [CrossRef] [PubMed]

[28]. McCulloch,W.S.; Pitts,W. A logical calculus of the ideas immanent in nervous activity. 1943. Bull. Math. Biol. 1990, 52, 99–115; discussion 73–97. [CrossRef]

[29]. Ahmed, S.S.S.J.; Santosh, W.; Kumar, S.; Christlet, H.T.T. Metabolic profiling of Parkinson's disease: Evidence of biomarker from gene expression analysis and rapid neural network detection. J. Biomed. Sci. 2009, 16, 63. [CrossRef] [PubMed]

[30]. O'Shea, K.; Cameron, S.J.S.; Lewis, K.E.; Lu, C.; Mur, L.A.J. Metabolomic-based biomarker discovery for non-invasive lung cancer screening: A case study. Biochim. Biophys. Acta BBA Lipids Lipid Metab. 2016, 1860, 2682–2687. [CrossRef] [PubMed]

[31]. Zhang, X.; Li, Y.; Liang, Y.; Sun, P.; Wu, X.; Song, J.; Sun, X.; Hong, M.; Gao, P.; Deng, D. Distinguishing Intracerebral Hemorrhage from Acute Cerebral Infarction through Metabolomics. Rev. Investig. Clin. Organo Hosp. Enferm. Nutr. 2017, 69, 319–328. [CrossRef] [PubMed]

[32]. Wei, Y.; Varanasi, R.S.; Schwarz, T.; Gomell, L.; Zhao, H.; Larson, D.J.; Sun, B.; Liu, G.; Chen, H.; Raabe, D.; et al. Machine learning-enhanced time-of-flight mass spectrometry analysis. Patterns 2021, 2, 100192. [CrossRef]

[33]. Heinemann, J. Machine Learning in Untargeted Metabolomics Experiments. Methods Mol. Biol. 2019, 1859, 287–299.

[34]. Le, T.T.; Fu, W.; Moore, J.H. Scaling tree-based automated machine learning to biomedical big data with a feature set selector.Bioinformatics 2020, 36, 250–256. [CrossRef]

[35]. Kuhn, M. Building predictive models in R using the caret package. J. Stat. Softw. 2008, 28, 1–26. [CrossRef]

[36]. Sen, P.; Lamichhane, S.; Mathema, V.B.; McGlinchey, A.; Dickens, A.M.; Khoomrung, S.; Orešiˇc, M. Deep learning meets

metabolomics: A methodological perspective. Brief. Bioinform. 2020, 22, 1531–1542. [CrossRef]

[37]. Casadei-Gardini, A.; Del Coco, L.;Marisi, G.; Conti, F.; Rovesti, G.; Ulivi, P.; Canale,M.; Frassineti, G.L.; Foschi, F.G.; Longo, S.; et al.1H-NMR Based SerumMetabolomics Highlights Different Specific Biomarkers between Early and Advanced Hepatocellular Carcinoma Stages. Cancers 2020, 12, 241. [CrossRef]

[38]. Manduchi, E.; Romano, J.D.; Moore, J.H. The promise of automated machine learning for the genetic analysis of complex traits. Hum. Genet. 2021, 141, 1529–1544. [CrossRef]

[39].. Liggi, S.; Hinz, C.; Hall, Z.; Santoru, M.L.; Poddighe, S.; Fjeldsted, J.; Atzori, L.; Griffin, J.L. KniMet: A pipeline for the processing of chromatography–mass spectrometry metabolomics data. Metabolomics 2018, 14, 52. [CrossRef]

[40]. Verhoeven, A.; Giera, M.; Mayboroda, O.A. KIMBLE: A versatile visual NMR metabolomics workbench in KNIME. Anal. Chim.Acta 2018, 1044, 66–76. [CrossRef]

[41]. Coelewij, L.;Waddington, K.E.; Robinson, G.A.; Chocano, E.;McDonnell, T.; Farinha, F.; Peng, J.; Dönnes, P.; Smith, E.; Croca, S.; et al. Using serummetabolomics analysis to predict sub-clinical atherosclerosis in patients with SLE. medRxiv 2020. [CrossRef]

[42]. Evans, E.D.; Duvallet, C.; Chu, N.D.; Oberst, M.K.; Murphy, M.A.; Rockafellow, I.; Sontag, D.; Alm, E.J. Predicting human health from biofluid-based metabolomics using machine learning. Sci. Rep. 2020, 10, 17635. [CrossRef] [PubMed]

[43]. Orlenko, A.; Kofink, D.; Lyytikäinen, L.P.; Nikus, K.; Mishra, P.; Kuukasjärvi, P.; Karhunen, P.J.; Kähönen, M.; Laurikka, J.O.; Lehtimäki, T.; et al. Model selection for metabolomics: Predicting diagnosis of coronary artery disease using automated machine learning. Bioinformatics 2020, 36, 1772–1778. [CrossRef] [PubMed]

[44]. Chen, H.; Wang, Z.; Qin, M.; Zhang, B.; Lin, L.; Ma, Q.; Liu, C.; Chen, X.; Li, H.; Lai, W.; et al. Comprehensive Metabolomics Identified the Prominent Role of Glycerophospholipid Metabolism in Coronary Artery Disease Progression. Front. Mol. Biosci. 2021, 8, 110. [CrossRef] [PubMed]

[45]. Wang, D.; Greenwood, P.; Klein, M.S. Deep Learning for Rapid Identification of Microbes Using Metabolomics Profiles. Metabolites 2021, 11, 863. [CrossRef] [PubMed]

[46]. Holzinger, A.; Biemann, C.; Pattichis, C.S.; Kell, D.B. What do we need to build explainable AI systems for the medical domain?arXiv 2017, arXiv:1712.0992