frontiers | Frontiers in Molecular Biosciences

TYPE Review
PUBLISHED 17 December 2024
DOI 10.3389/fmolb.2024.1483326

Check for updates

OPEN ACCESS

EDITED BY
Michele Costanzo,
University of Naples Federico II, Italy

REVIEWED BY
Giuseppina Fanelli,
University of Tuscia, Italy
Renu Pandey,
Indian Institute of Technology Bombay, India

*CORRESPONDENCE
Zhibin Zhang,
zhibin-zhang@nankai.edu.cn
Xiangyang Zhang,
xiangyang.zhang@tju.edu.cn

RECEIVED 23 August 2024
ACCEPTED 02 December 2024
PUBLISHED 17 December 2024

CITATION
Che Y, Zhao M, Gao Y, Zhang Z and Zhang X (2024) Application of machine learning for mass spectrometry-based multi-omics in thyroid diseases.
Front. Mol. Biosci. 11:1483326.
doi: 10.3389/fmolb.2024.1483326

COPYRIGHT
© 2024 Che, Zhao, Gao, Zhang and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Application of machine learning for mass spectrometry-based multi-omics in thyroid diseases

Yanan Che[1], Meng Zhao[1], Yan Gao[1], Zhibin Zhang[2]* and Xiangyang Zhang[1]*

[1]School of Pharmaceutical Science and Technology, Tianjin University, Tianjin, China, [2]Department of General Surgery, Tianjin First Central Hospital, Tianjin, China

Thyroid diseases, including functional and neoplastic diseases, bring a huge burden to people's health. Therefore, a timely and accurate diagnosis is necessary. Mass spectrometry (MS) based multi-omics has become an effective strategy to reveal the complex biological mechanisms of thyroid diseases. The exponential growth of biomedical data has promoted the applications of machine learning (ML) techniques to address new challenges in biology and clinical research. In this review, we presented the detailed review of applications of ML for MS-based multi-omics in thyroid disease. It is primarily divided into two sections. In the first section, MS-based multi-omics, primarily proteomics and metabolomics, and their applications in clinical diseases are briefly discussed. In the second section, several commonly used unsupervised learning and supervised algorithms, such as principal component analysis, hierarchical clustering, random forest, and support vector machines are addressed, and the integration of ML techniques with MS-based multi-omics data and its application in thyroid disease diagnosis is explored.

KEYWORDS

mass spectrometry, proteomics, metabolomics, multi-omics, thyroid diseases, machine learning

1 Introduction

The thyroid gland is a small, butterfly-shaped gland located at the base of the neck (Sofia et al., 2019; Mullur et al., 2014). It plays a crucial role in regulating various metabolic processes by secreting hormones (Sofia et al., 2019; Mullur et al., 2014). Thyroid disease refers to various diseases affecting the thyroid gland, categorized into functional and neoplastic diseases (Vanderpump, 2011; Zhang et al., 2022). Functional diseases are classified as hyperthyroidism or hypothyroidism, whereas neoplastic diseases are classified as benign or malignant (Zhang et al., 2022).

In the field of neoplastic diseases, tumors are classified as benign tumors, low-risk neoplasms, and malignant neoplasms according to prognostic risk categories (Basolo et al., 2023). Thyroid cancer refers to malignant tumors, originating from follicular or parafollicular thyroid cells, which can metastasize to other places in the body (Omur and Baran, 2014). Thyroid cancer is one of the most common endocrine neoplasia, and its incidence has been on the rise in the past 40 years, disproportionately affecting women (Chen et al., 2023a; Guarino et al., 2010).

According to "The 5th edition of the World Health Organization (WHO) classification of endocrine tumors" which was released in 2022, thyroid cancer exists in several

forms (Schneider and Chen, 2013), including differentiated thyroid cancer (DTC), undifferentiated thyroid cancer, and medullary thyroid cancer (MTC). DTC, the most prevalent type of thyroid malignancy, primarily includes papillary thyroid carcinoma (PTC), follicular thyroid carcinoma (FTC), and oncocytic thyroid carcinoma, with PTC accounting for 85% –90% of all DTC cases (Omur and Baran, 2014; Caria et al., 2019). Thyroid cancer presents a complex and clinically significant challenge. To explore the molecular mechanisms of thyroid cancer, researchers have increasingly turned to omics approaches.

Omics is a technique for the comprehensive evaluation of different classes of biomolecules, including genomics, transcriptomics, proteomics, metabolomics, and others (Babu and Snyder, 2023). Using only one type of data to understand the characteristics and complications of a disease is not enough. Recently, exhaustive exploration through multi-omics strategies has garnered increasing attention among analytical chemists (Kappler and Lehmann, 2019). Advances in various omics technologies, such as proteomics and metabolomics, coupled with enhanced computing capabilities, have paved the way for innovative integration of diverse omics data (Babu and Snyder, 2023). With the rapid development of high throughput sequencing and multi-omics, biomedical research has increasingly adopted a combination of multi-omics technologies. Multi-omics strategies aim to scrutinize the same samples using two or more omics methods, integrating diverse omics data to reveal coherent associations and attain a comprehensive, holistic understanding of biomedical processes (Kappler and Lehmann, 2019).

Mass spectrometry (MS) is crucial for studying multi-omics. It is a high-throughput analytical technology that can quantify countless molecules, from metabolites and lipids to peptides and proteins (Zhao et al., 2022a; Leung Kwan et al., 2021). This analytical technology aids in discovering biomarkers, understanding diseases at the molecular level, and provides a new perspective in the biological field (Leung Kwan et al., 2021). As an emerging approach of biomarker discovery, MS-based multi-omics plays a significant role in the early diagnosis and screening, classification, and prognosis of diseases. However, the large amounts of data generated by high-throughput technologies require specialized data analysis strategies (Zhao et al., 2022b).

Machine learning (ML) is a driving force behind data integration in systems biology (Alber et al., 2019). Through data-driven bioinformatics analysis of MS-based multi-omics data, ML serves as a powerful tool for revealing the intrinsic mechanisms of various biological events (Leung Kwan et al., 2021). The combination of MS-based multi-omics and advanced data integration approach holds promise for deeper investigation of complex biological processes.

In this review, we presented the detailed review of applications of ML for MS-based multi-omics in thyroid disease. In literature previously published, applications of ML in thyroid disease or applications of MS-based multi-omics in thyroid disease are reviewed, but no one has combined them into a comprehensive review. This review can provide new insights to the people who focuses on applications of combining ML with MS-based multi-omics in thyroid disease. It is primarily divided into two sections. The first section briefly introduces MS-based multi-omics, mainly proteomics and metabolomics, and their applications in clinical diseases. The second section addresses a comprehensive overview of ML models, and explores the integration of ML techniques into MS-based multi-omics data, and its application in thyroid disease diagnosis.

2 Mass spectrometry-based multi-omics in thyroid diseases

Data from various studies, including genomics, transcriptomics, proteomics, and metabolomics studies together are denoted as "multi-omics" data (Figure 1). Individual datasets from these "-omics" studies can serve as valuable biomarkers for studying, exploring, and understanding the traits and complexities of biological organisms (Manochkumar et al., 2023).

MS plays a crucial role in multi-omics research by detecting metabolites or proteins in samples (Qiu et al., 2023). The use of mass spectrometry technology for detecting metabolites or proteins in samples can identify thousands of proteins or metabolites across a substantial volume of samples (Kowalczyk et al., 2020). This high-throughput approach not only improves our ability to identify molecular signatures but also helps us gain a more comprehensive understanding of the intricate biological processes within organisms (Leung Kwan et al., 2021).

2.1 Mass spectrometry-based proteomics

Oncogenesis is associated with changes in the levels of various proteins involved in cell proliferation, migration, and apoptosis (Migisha et al., 2020). Proteomics enables the maximum identification and quantification of all proteins in cells or tissues, establishes the connection between genes and their corresponding protein products, and provides information about proteins, including their subcellular localization, post-translational modifications, and interactions with other proteins, aiming to reveal the mechanisms behind their biological functions (Manochkumar et al., 2023; Chen et al., 2023b; Kang et al., 2022). The analysis of the proteome can provide valuable insights into the fundamental molecular mechanisms of diseases, responses to therapy, and the identification of diagnostic, predictive biomarkers and prognostic crucial for precision medicine (Ball et al., 2023).

Mass spectrometry (MS) is an analytical technique that measures the mass-to-charge ratio (m/z) of ionized molecules. The basic components of a mass spectrometer include the ion source, mass analyzer, and detector. Proteins or peptides are ionized in the ion source, separated based on their m/z in the mass analyzer, and detected to generate a mass spectrum. This mass spectrum provides detailed information about the molecular weight and structural characteristics of the analyte. Mass spectrometry-based proteomics mainly includes five processes (Figure 2).

2.1.1 Sample preparation

The proteomics workflow begins with the preparation of biological samples. Proteins are extracted from the sample, often followed by enrichment or fractionation to reduce complexity. This step is critical for ensuring the accurate identification and quantification of proteins.

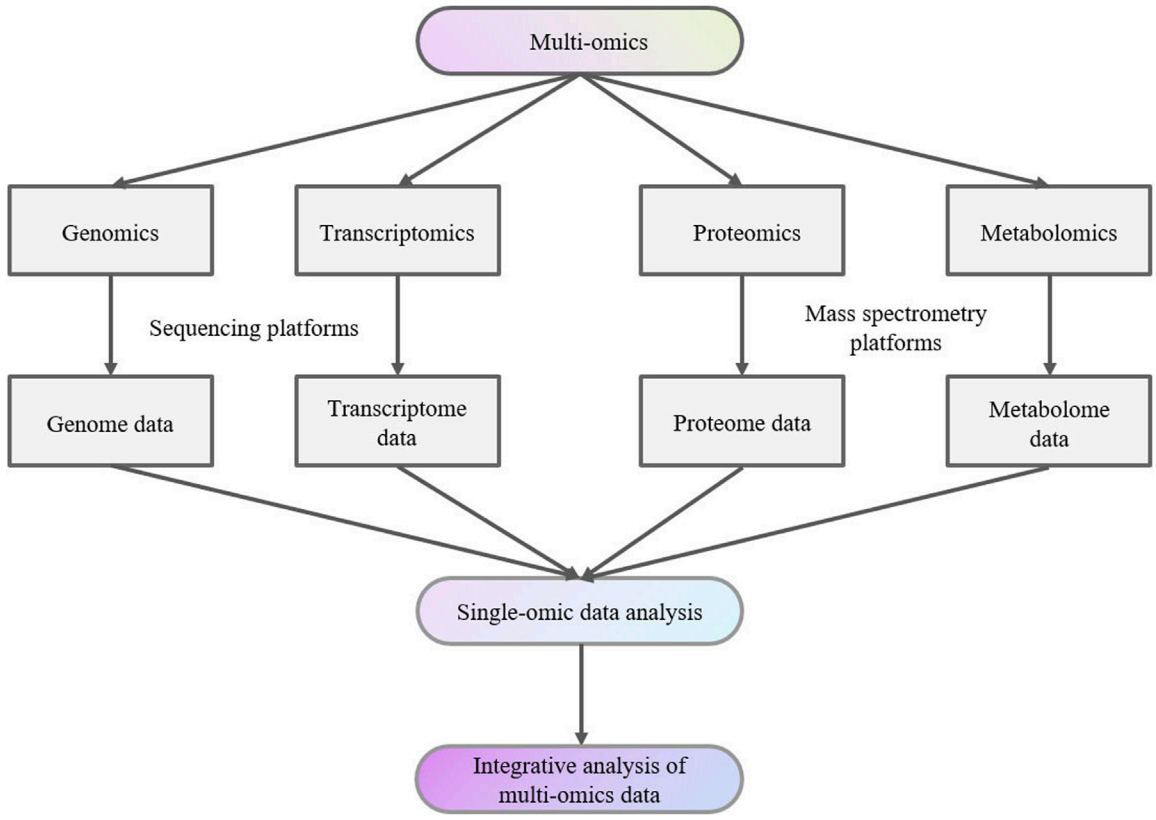

FIGURE 1
Multi-omics analysis flow chart.

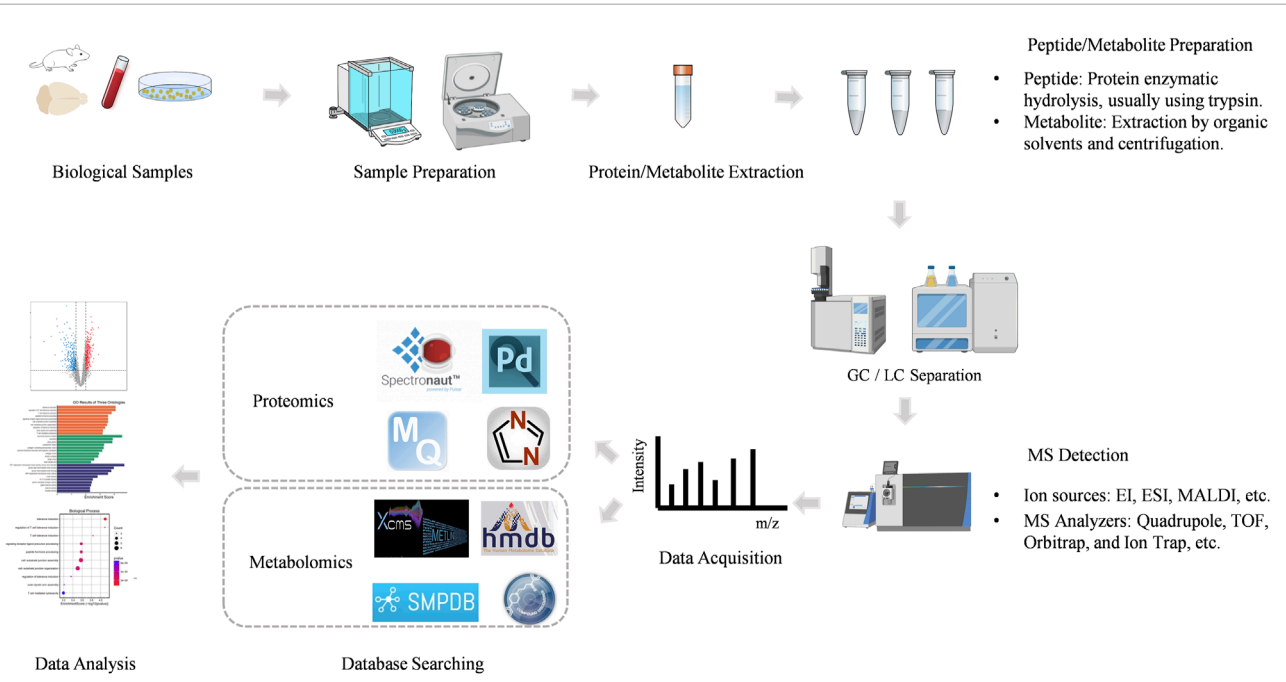

FIGURE 2
The workflow of MS-based proteomics and metabolomics analysis.

The proteins are then digested into smaller peptides, typically using an enzyme like trypsin. This peptide mixture is more amenable to analysis by MS.

2.1.2 Peptide ionization

Peptides are ionized in the ion source, which can be achieved through various techniques. The most common ionization methods in proteomics are Electrospray Ionization (ESI) and Matrix-Assisted Laser Desorption/Ionization (MALDI).

ESI is widely used for liquid chromatography-mass spectrometry (LC-MS) and generates ions by applying a high voltage to a liquid sample, producing charged droplets that release ions. MALDI involves embedding the sample in a matrix that absorbs laser energy, leading to the ionization of peptides.

2.1.3 Mass analysis

The ionized peptides are introduced into the mass analyzer, where they are separated based on their m/z ratio. There are several types of mass analyzers, including Quadrupole, Time-of-Flight (TOF), Orbitrap, and Ion Trap, each offering different advantages in terms of resolution, accuracy, and speed.

High-resolution mass analyzers, such as the Orbitrap and TOF, are particularly valuable in proteomics for their ability to distinguish between ions with very similar m/z ratios, enabling the identification of complex peptide mixtures.

2.1.4 Peptide fragmentation

To obtain sequence information, peptides are often subjected to fragmentation in the mass spectrometer. Tandem mass spectrometry (MS/MS) involves two stages of mass analysis: the first stage selects a precursor ion, which is then fragmented, and the second stage analyzes the resulting fragment ions.

The fragmentation patterns are characteristic of the peptide's amino acid sequence, allowing for the identification of the peptide and inference of the protein from which it originated (Searle et al., 2020).

2.1.5 Data analysis

The mass spectrometry data are processed using sophisticated bioinformatics tools. Software such as MaxQuant, Proteome Discoverer, DIA-NN, Peaks, Spectronaut and Mascot matches the obtained mass spectra to theoretical spectra derived from protein databases, enabling protein identification.

Quantitative proteomics can be achieved through various techniques, including label-free quantification, stable isotope labeling (e.g., SILAC, iTRAQ), and tandem mass tags (TMT). These approaches allow for the relative or absolute quantification of proteins across different samples.

The development of highly sensitive and high-throughput MS platforms over the past decade means that it is now possible to identify and quantify thousands of proteins from large numbers of biological samples. Rapid advancements in MS and data analysis strategies have significantly enhanced proteomics research worldwide (Halder et al., 2021). MS-based proteomics is increasingly recognized as a widely adopted technique for characterizing proteomes (Migisha et al., 2020). Proteomics research can be divided into untargeted proteomics and targeted proteomics. Untargeted proteomics is also called discovery proteomics, which detects differential proteins in different samples by detecting proteins as many as possible. The research objects of untargeted proteomics are uncertain, and are often all the protein or peptide components contained in the sample, which are relatively large in number. Targeted proteomics is the quantitative detection of target proteins. The research objects of targeted proteomics are specific and the number is relatively small. Compared with untargeted proteomics, it has greater sensitivity and accuracy and is often used for verification analysis of biomarkers. Data-dependent acquisition (DDA) and data-independent acquisition (DIA) are the two primary MS strategies for untargeted proteomics (Qian et al., 2023). DDA is a traditional MS-based proteomics analysis method. In DDA, in the second stage of tandem mass spectrometry, a small number of peptides are selected for fragmentation within a narrow range of mass-to-charge ratio (m/z) signal intensity (Hu et al., 2016). DIA is another MS-based proteomics analysis method. DIA divides the entire full scan range of the mass spectrometer into several windows and then fragments all peptide precursors within each window simultaneously to generate a comprehensive MS^2 spectrum (Kawashima et al., 2019; Wang et al., 2022a).

2.2 Mass spectrometry-based metabolomics

Metabolomics is to detect and qualitatively and quantitatively analyze the dynamic changes of metabolites of organisms, tissues or cells before and after a specific stimulus or interference which was initially introduced in 1999 by Jeremy-Nicholson and is an emerging research field (Nicholson et al., 1999; Wang et al., 2023a). The research objects are metabolites, which are mostly small molecule substances with a molecular mass range of ≤1,000 Da, such as small organics: acids, amino acids, nucleotides, sugars, lipids, vitamins, etc. Metabolites are the end products of cellular processes and can directly reflect the physiological state of an organism. Liquid chromatography coupled to mass spectrometry (LC-MS) was first used to study thyroid cancer in serum samples in 2011 (Yao et al., 2011). DIA workflow was applied for metabolomics in 2017 (Zhou et al., 2017).

MS-based metabolomics involves the separation, detection, and characterization of metabolites, providing comprehensive coverage of the metabolome. The workflow is shown in Figure 2.

2.2.1 Sample preparation

The first step in MS-based metabolomics involves the preparation of biological samples. Metabolites can be extracted from various biological matrices, such as plasma, urine, tissues, or cell cultures, using extraction methods optimized for different classes of metabolites.

Sample preparation is critical to preserving the integrity of the metabolome and avoiding contamination or degradation. The extracted metabolites are often subjected to concentrate to improve the detection of low-abundance compounds.

2.2.2 Metabolite separation

Prior to mass spectrometric analysis, metabolites are typically separated using chromatographic techniques to reduce sample

complexity. The most common techniques are Gas Chromatography (GC) and Liquid Chromatography (LC).

Gas Chromatography-Mass Spectrometry (GC-MS) is particularly well-suited for analyzing volatile and semi-volatile compounds. In GC-MS, metabolites are vaporized and separated in a gas phase before being ionized and detected by a mass spectrometer.

Liquid Chromatography-Mass Spectrometry (LC-MS) is more versatile and can handle a broader range of metabolites, including polar, non-volatile, and thermally labile compounds. LC-MS separates metabolites in a liquid phase based on their interaction with the stationary phase and then ionizes them for mass spectrometric detection.

2.2.3 Ionization of metabolites

The ionization of metabolites is a crucial step in mass spectrometry, as it converts neutral molecules into charged ions that can be detected. Common ionization methods include Electrospray Ionization (ESI) and Atmospheric Pressure Chemical Ionization (APCI) for LC-MS, and Electron Ionization (EI) for GC-MS.

ESI is widely used in LC-MS due to its ability to ionize a wide range of metabolites, particularly those that are polar and easily ionizable. ESI produces ions by applying a high voltage to the liquid sample, resulting in charged droplets that release ions as they evaporate.

EI, commonly used in GC-MS, involves bombarding gas-phase molecules with high-energy electrons, leading to ionization and fragmentation. The resulting fragment ions provide structural information about the metabolite.

2.2.4 Mass analysis and detection

Once ionized, metabolites are introduced into the mass analyzer, where they are separated based on their m/z ratio. Unlike proteomics, metabolomics is divided into positive and negative ion modes due to the different properties of the compounds. Various mass analyzers are used in metabolomics, including Quadrupole, Time-of-Flight (TOF), Orbitrap, and Ion Trap analyzers.

High-resolution mass analyzers, such as the Orbitrap and TOF, are particularly valuable in metabolomics for their ability to accurately measure the m/z of metabolites and distinguish between compounds with very similar masses.

2.2.5 Data acquisition and processing

The mass spectrometer generates a mass spectrum, which provides information on the m/z ratios and intensities of detected ions. This data is then processed using specialized software to identify and quantify metabolites.

The identification of metabolites is typically performed by matching the acquired mass spectra against reference libraries, databases or, such as HMDB, Compound Discover, METLIN and SMPDB (Xiao et al., 2012). Accurate mass measurements and fragmentation patterns are used to deduce the molecular structure of unknown metabolites.

Metabolites can be quantified either relatively, by comparing the intensity of ion signals between samples, or absolutely, using calibration curves with known standards.

MS is a major platform for clinical metabolomics due to its excellent sensitivity, selectivity, and wide dynamic range (Ding and Feng, 2023). MS-based metabolomics can simultaneously detect and quantify thousands of metabolite features (Alseekh et al., 2021). Common MS-based metabolomics methods include GC-MS and LC-MS. Compared with GC-MS, LC-MS generates extensive data, has high sensitivity, and can measure a wide range of metabolites. Due to the feasibility of liquid chromatography (LC) in separating a wide range of metabolites with broad polarity, combining LC with high-resolution MS systems consistently detects and quantifies thousands of metabolic features, even from minimal sample amounts such as 10 mg of tissue, 50 uL of urine, or as few as half a million cells (Guo et al., 2022a). LC-MS-based metabolomics has gained increasing attention for identifying disease biomarkers and providing unique insights into pathophysiological processes (Ding and Feng, 2023; Randall et al., 2023).

Mass spectrometry imaging (MSI) technology is also widely used in the study of the spatiotemporal distribution of various metabolites, peptides and proteins in animal/plant tissues due to its advantages such as label-free, non-specific, high sensitivity, high chemical coverage, and simultaneous detection of elements/molecules.

2.3 Mass spectrometry-based multi-omics applications in thyroid diseases

High-throughput techniques, exemplified by MS, play a crucial role in the measurement of metabolomic and proteomic data (Reel et al., 2021). Collectively, these "-omics" data hold the potential to significantly advance precision medicine, particularly in the context of biomarker-driven approaches for conditions such as endocrine diseases, diabetes, cancer, cardiovascular disease, respiratory disorders, and Alzheimer's disease (Reel et al., 2021).

Regarding thyroid diseases, understanding the pathogenesis is essential for improving diagnostic accuracy, precise risk stratification, and enabling personalized treatment (Li et al., 2023a). In recent years, with the continuous development of MS, various omics analysis methods based on different sample types (cells, tissues, serum, and urine) have been applied to the study of thyroid disease, actively promoting the development of accurate diagnosis and treatment of thyroid disease by clarifying the pathogenesis, diagnostic grading, prognosis prediction and targeted therapy (Li et al., 2023a).

Biomarkers refer to "an indicator that can be objectively detected and evaluated and can be used as an indicator of normal biological processes, pathological processes, or pharmacological responses to therapeutic intervention" and are of great significance for screening, diagnosing, or monitoring diseases (Biomarkers, 2001; Mischak et al., 2010; Joshi et al., 2024). The exploration of biomarker discovery holds promise in identifying potential markers for early disease detection, prognosis assessment, predicting and monitoring treatment responses (Jimenez and Verheul, 2014). The identification and validation of reliable biomarkers will continue to help improve our understanding of thyroid disease and refine treatment strategies (Davis et al., 2020; Califf, 2018).

Misdiagnosis is common in the diagnosis of thyroid disease (Walsh, 2016). Therefore, it is necessary to identify biomarkers for specific thyroid disease states. MS-based proteomics and metabolomics have been widely used for the discovery of potential biomarkers in the research of thyroid disease.

Much of the published proteomic studies of thyroid disease have compared the protein profiles of thyroid disease groups with healthy thyroid groups to find potential protein markers (Paron et al., 2003). Tissues and cell lines of thyroid are always used for differential proteomics. In a 1997 study, Galectin-3 was proposed to be a potential biomarker of malignant thyroid tumors, especially papillary carcinomas (Fernández et al., 1997). This finding has been confirmed by several other independent researches using different proteomic approaches: MALDI-MSI (Paron et al., 2003), two-dimensional gel electrophoresis and LC-MS (Torres-Cabala et al., 2004). S100 family proteins are comprised of 21 small isoforms, and many of them implicated in important cellular functions such as proliferation, motility and survival (Martinez-Aguilar et al., 2015). Several papers have been published confirming them as potential biomarkers in thyroid cancer by proteomic approaches (Torres-Cabala et al., 2004; Nishimura et al., 1997; Wang et al., 2021). Torres-Cabala, C. et al. identified a new protein, S100C, which is highly expressed in PTC by two-dimensional gel electrophoresis and LC-MS (Torres-Cabala et al., 2004). S100A6 was found to be expressed at a significantly higher level in PTC compared with other tumor groups or normal tissues by LC-MS based proteomics (Sofiadis et al., 2010). Nipp et al. (2012) confirmed S100A10 and S100A6 as biomarkers of PTC with lymph node metastasis identified by MALDI-MSI proteomic approach. This result also demonstrated the potential application of MALDI-MSI proteomic approach in identifying biomarkers in thyroid cancer.

In recent years, exosomes, small membrane microvesicles derived from endosomal cells, have attracted great interest in the proteomics of thyroid diseases due to their role in transporting proteins, lipids and nucleic acids into target cells (Zhang et al., 2019). Transport of molecules via exosomes is one of the factors in the development of thyroid cancer, and the transported molecules can serve as cancer biomarkers (Surman et al., 2024). Luo et al. (2018) compared proteome profiles of serum-purified exosomes (SPEs) from PTC patients with LNM, PTC patients without LNM, and healthy donors. The results showed that specific proteins related to cancer cell metastasis, such as SRC, TLN1, ITGB2, and CAPNS1, were overexpressed in the SPEs of PTC patients with LNM (Luo et al., 2018). In the study of Xi Jia et al., the screened differentially expressed proteins, such as MAP1S, VAMP8, IF5, RSU1, ACTB and CXCL7, were mainly enriched in the immune system and metabolic system that can be seen as potential biomarkers, indicating that plasma exosomes may play an important role in the systemic immune imbalance of autoimmune thyroid diseases (AITDs) (Jia et al., 2021).

Proteomics can not only provide biomarkers for diagnosis but also reveal potential therapeutic targets. For example, protein HSP90 was found to be overexpressed in thyroid cancer (Pearl et al., 2008; Liu et al., 2017). HSP90 regulates protein degradation of several growth-mediating kinases such as BRAF and RET which are well known for the role they play in carcinogenesis (Gild et al., 2016). Several studies have shown that inhibition of HSP90 can not only attenuate cell proliferation but also improve the efficacy of radioiodine therapy in thyroid cancer patients (Gild et al., 2016; Marsee et al., 2004; Wickenberg et al., 2024; White et al., 2016).

Since tumors significantly alter major metabolic pathways, metabolomics is also rapidly becoming an important method for identifying cancer biomarkers. Alterations of the metabolome can be reflected in both tissues and biological fluids. Most chromatography-based metabolomics studies focus on biomarkers between disease and normal groups. Huang et al. (2019a) conducted metabolomic studies using 1,540 clinical serum and plasma samples, along with 114 clinical tissue samples, to characterize the metabolomic profiles of healthy controls and patients with thyroid nodules, including benign thyroid nodules (BTN) and PTC. Their research identified a group of circulating metabolites—myo-inositol, α-N-phenylacetyl-L-glutamine, proline betaine, L-glutamic acid, LysoPC (18:0), and LysoPC (18:1)—as potential biomarkers. Jajin et al. (2022) used GC-MS to perform plasma metabolomics profiling of medullary thyroid cancer (MTC) patients. Results showed that linoleic acid, linolenic acid, and leucine can be used as potential biomarkers for early detection of MTC. These findings provide a basis for the diagnosis and management of thyroid cancer patients from a metabolomics perspective.

Spatially resolved metabolomics integrates MSI and metabolomics technology to accurately measure the types, contents and differential spatial distribution of endogenous or exogenous metabolites in biological tissues and cells and shows great prospect in biomarker discovery of thyroid disease. Jialing Zhang et al. (2017b) used desorption electrospray ionization mass spectrometry imaging (DESI-MSI) to analyze metastatic thyroid cancer in human lymph node tissues and the results showed that the relative abundance of ceramide and glycerophosphoinositide increased.

Wojakowska et al. (2018) used MALDI-MSI to analysis of lipid distribution directly in formalin-fixed tissue. The results showed that the abundance of phosphatidylcholine (32:0, 32:1, 34:1 and 36:3), sphingomyelin (34:1 and 36:1) and phosphatidic acid (36:2 and 36:3) were significantly higher in cancer tissues than them in non-cancer tissues (Wojakowska et al., 2018).

Luojiao Huang et al. (2019b) used the air-flow assisted desorption electrospray ionization (AFADESI) MSI to investigated the metabolic characteristics of different microregions of PTC and results showed that phenylalanine, leucine and tyrosine were expressed at the highest levels in tumors, with a trend of gradually decreasing from tumors to stromal tissues and normal tissues, while creatinine was the opposite.

Biomarker discovery can contribute to molecular subtyping in thyroid disease. The integration of MS-based multi-omics is a powerful tool for elucidating complex molecular signatures of various cancer subtypes (Wang et al., 2023b). This approach not only enhances our understanding of the mechanisms of action of various molecules within cancer but also facilitates more targeted and personalized therapeutic interventions for specific subtypes (Berger and Mardis, 2018).

Martinez-Aguilar et al. (2016) applied DIA MS for quantitative analysis of expression levels for over 1,600 proteins across 32 specimens, discerning differences between normal thyroid tissue and the three prevalent thyroid gland tumors: follicular adenoma, follicular carcinoma, and papillary carcinoma. Proteomic pathway analysis revealed that changes in papillary carcinomas are associated with disruption of cell contacts (loss of E-cadherin), actin cytoskeletal dynamics, and loss of differentiation markers, characteristics of the aggressive phenotype (Martinez-Aguilar et al., 2016).

Wojakowska et al. (2015) used the GC-MS method to extract, identify, and semi-quantitate metabolites in formalin-fixed

paraffin-embedded (FFPE) tissue specimens from five different types of thyroid malignancies, benign follicular adenoma and normal thyroid and concluded that multicomponent metabolomic signatures can be used to classify different subtypes of follicular thyroid lesions.

MS-based multi-omics have significantly increased in recent years and enabled mapping of biochemical changes in thyroid disease and hence can provide an opportunity to develop predictive biomarkers that can trigger earlier interventions (Wang et al., 2013). MS can directly quantify thyroid analytes and its high resolution can enhance the accuracy and detection (Jasem et al., 2024). Biomarkers of thyroid disease screened out by MS-based proteomics and metabolomics can not only provide a basis for clinical diagnosis, but also provide insights into the biological mechanisms of thyroid disease. It can be used to distinguish different types of thyroid cancer, which is beneficial for classifying benign and malignant cancers for treatment, such as using different dosing strategies, thereby achieving precision medicine.

3 Classic machine learning models and multi-omics applications in thyroid disease

This section will present several classic machine learning models and offer an overview of how these models contribute to the study of thyroid diseases.

3.1 Classic machine learning models in data analysis

Machine learning (ML) is the science of developing algorithms and statistical models, which is a subset of artificial intelligence (Smith et al., 2023). Computer systems utilize these algorithms and models to perform tasks without explicit instructions, enabling machines to undertake activities requiring human intelligence, such as diagnosis, planning, and prediction, based on established patterns and reasoning (Mohammadzadeh et al., 2024). In recent years, the exponential growth of biomedical data has driven many applications of ML techniques to address new challenges in biology and clinical research (Auslander et al., 2021). ML methods are favored in statistical analysis because of their inherent nonlinear data representation and ability to quickly process large datasets (Liebal et al., 2020). ML algorithms are employed for training, key feature identification, and group classification (Huang et al., 2018).

Generally, ML can be categorized into four main types: unsupervised, supervised, semi-supervised, and reinforcement learning (Figure 3). Current research in clinical diseases predominantly focuses on unsupervised and supervised learning algorithms, which will be the focus of this review (Perakakis et al., 2018).

This section examines the role of ML in handling and analyzing the vast and complex datasets generated by MS-based multi-omics approaches. It discusses specific algorithms and techniques for data processing, feature selection, and classification, emphasizing their importance in identifying potential biomarkers and therapeutic targets.

3.1.1 Unsupervised learning

Unsupervised learning involves using datasets that contain only input data and attempts to find structure in the data by grouping or clustering the data points (Angra and Ahuja, 2017). Unsupervised learning algorithms are primarily employed for dimensionality reduction, clustering, and association tasks (Arjmand et al., 2022). Four common unsupervised algorithms include principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), uniform manifold approximation and projection (UMAP), and hierarchical clustering (HCL).

3.1.1.1 Principal component analysis (PCA)

The complexity of multivariate data often necessitates the use of dimensionality reduction methods to simplify the information. Dimensionality reduction of high-dimensional data plays a crucial role in downstream tasks such as pattern recognition, classification, and clustering (Kim et al., 2018). Principal component analysis (PCA) is a classic unsupervised dimensionality reduction method that identifies hidden features in data, providing the most significant signals, and is often used in machine learning. (Kim et al., 2018; Chen and Gao, 2016; Ma and Dai, 2011). PCA simplifies complex data and makes the analysis process easier. Essentially, PCA is an "unsupervised" method that analyzes data purely based on its characteristics, without knowing the grouping of each sample. PCA effectively identifies the "main" elements and structures in the data, removes noise and redundancy, reduces the dimensionality of complex data, and reveals the simple structure hidden behind the complex data (Sugimoto et al., 2012). PCA is widely used in MS-based multi-omics data analysis, particularly for data dimensionality and achieving data visualization (Sugimoto et al., 2012). The results are often visualized using scatter plots.

PCA is commonly applied in clinical analysis to reveal differences between samples, with the distance between samples on the horizontal and vertical axes representing the similarity distance of the samples under the influence of the principal components (PC1 and PC2).

For example, in a single-cell proteomics study of hepatocytes by Rosenberger et al. (2023), PCA was used to reduce the dimensionality of proteomics data, resulting in clear hepatocyte partitioning and demonstrating the biological validity of the data. Similarly, Xu et al. (2022) applied PCA to validate the metabolic profile of mouse liver tissue in a study on the mechanism of action of Huang Qin decoction for treating diabetic liver injury.

PCA is particularly suitable for initial exploratory data analysis, especially when linear relationships are presumed in the data and the interpretability of components is crucial (Beattie and Esmonde-White, 2021; Ivosev et al., 2008). It is also preferred in scenarios where computational efficiency is a priority. However, PCA may not adequately capture complex nonlinear interactions present in biological data.

3.1.1.2 t-distributed stochastic neighbor embedding (t-SNE)

t-distributed Stochastic Neighbor Embedding (t-SNE) is a nonlinear dimensionality reduction technique primarily designed for data visualization and excels at identifying and discovering complex nonlinear structures in data (Cieslak et al., 2020). It

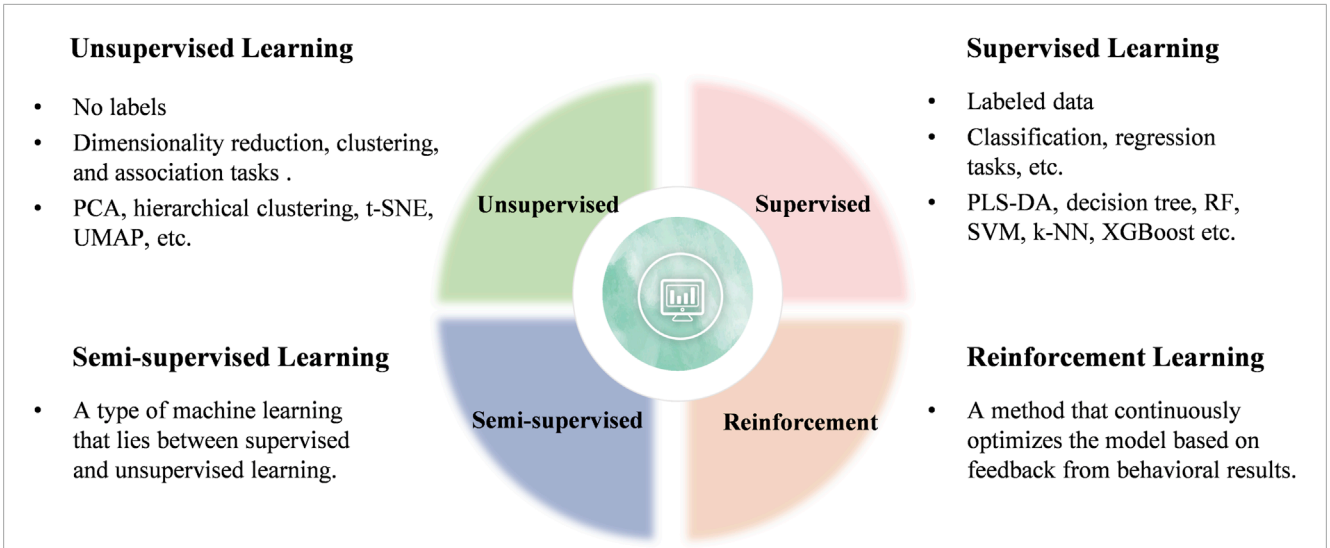

FIGURE 3
Classification of ML learning algorithms.

converts high-dimensional data into a low-dimensional space, typically two or three dimensions while preserving the local structure of the data (Cieslak et al., 2020; Da Silva Lopes et al., 2020). t-SNE focuses on preserving the relative distances between similar data points, making it effective for revealing clusters and patterns (Chatzimparmpas et al., 2020).

In the study by Liang et al. (2022), 1,681 proteins were analyzed through proteomics in 258 HCM patients. t-SNE was utilized to visualize and reduce the dimensionality of the data, revealing that the four molecular subtypes were well separated.

t-SNE is ideal for visualizing small to medium-sized datasets where the primary goal is to understand and explore local structures and clustering. It is particularly suitable for high-dimensional data with complex, nonlinear relationships (Gisbrecht et al., 2015). However, it is relatively slow, especially when applied to large datasets.

3.1.1.3 Uniform manifold approximation and projection (UMAP)

Uniform manifold approximation and projection (UMAP), an algorithm developed by McInnes et al., is a nonlinear dimensionality reduction technique (Lundberg and Lee, 2017). It is another nonlinear dimensionality reduction method particularly effective at preserving both local and some global structures of the data (Becht et al., 2018; Yang et al., 2021). UMAP is based on manifold learning techniques and constructs a high-dimensional graph representation of the data, which is subsequently optimized to create a low-dimensional embedding (McInnes et al., 2018).

Like t-SNE, UMAP is also effective in capturing complex nonlinear relationships in the data. In general, UMAP is faster and more scalable than t-SNE and it better preserves both local and global structures.

UMAP is preferred for large datasets requiring a balance between local and global structure preservation (Sainburg et al., 2021). It is effective for visualizing complex data structures and works well for high-dimensional data (Becht et al., 2018). Compared to PCA, this method is more complex and can be challenging to understand and debug. Although UMAP is faster than t-SNE, it still requires significant computational resources for very large datasets (Roca et al., 2023).

3.1.1.4 Hierarchical clustering (HCL)

A crucial component of unsupervised learning is the clustering algorithm. Traditionally, cluster analysis is classified as unsupervised learning because it does not involve class labels or quantitative response variables, which are characteristic of supervised learning methods such as classification and regression (Pan et al., 2013). Cluster analysis is a series of different algorithms that divide observation data into different categories or clusters based on distance functions (Blekherman et al., 2011). The goal is to partition the data into groups such that the distance between samples within each group is smaller than the distance between samples in different groups (Blekherman et al., 2011). Hierarchical clustering (HCL) is a clustering method frequently used in marker screening and enables visualization of gene, protein, and metabolite features (Picard et al., 2023; Granato et al., 2018). The results of HCL are commonly visualized using heatmaps.

In biomedical informatics, HCL is often applied to cluster protein sequence data. Proteins with similar structures also have similar functions. Proteins with similar functions can be grouped into categories through clustering, aiding in the study of protein functions. In clinical analysis, hierarchical clustering is utilized to intuitively display relationships between groups and highlights expression differences of characteristic substances.

In the study of MS-based urine proteomics of gastric lesions by Fan et al. (2022), HCL was used to partition 139 differential proteins with VIP>1 into six clusters, revealing dynamic changes from precancerous lesions dynamic changes in gastric cancer.

In an MS-based metabolomics study of cancer cell lines by Li et al. (2019), HCL was employed to assess metabolic similarities between cell lines.

For MS-based spatial proteomics, dimensionality reduction and clustering methods such as PCA, t-SNE and HCL are effective for quality control of MS-based spatial proteomic data and for examining organelle separation (Huang et al., 2022; Karimpour-Fard et al., 2015; Mou et al., 2022; Ringner, 2008).

3.1.2 Supervised learning

Supervised learning relies on labeled datasets to train algorithms on a predefined classification system and to infer the functional relationship between input features and output labels based on this training. The algorithm learns a function from the training dataset that enables prediction of outcomes for new data. In essence, supervised learning involves determining whether the objective is to predict outcomes based on known input-output pairs. A significant category within supervised learning is classification problems. In the classification problems, the target variables are discrete rather than continuous. Examples include tumor size, patient age, and the benign or malignant status of the tumor.

The process of model training with ML algorithms involves three steps: data splitting, parameter estimation using the training set, and performance evaluation using the test set.

a. Data Splitting

Typically, the dataset used to train a ML model is divided into a training set and a test set, with a common ratio of 70:30 (Galal et al., 2022). A validation set is often included for model performance evaluation and hyperparameter tuning, ensuring optimal results under the given data conditions. In this scenario, the data can be divided into 60% training, 20% validation, and 20% test sets (Galal et al., 2022).

b. Parameter Estimation Using Training Data

Parameter estimation is a critical step in model training. The goal of using training set to estimate model parameters is to create a model that accurately captures the underlying information in the data so that it can make reliable predictions about new data.

c. Comprehensive Evaluation Using Test Data

The test set evaluates the overall performance of the final model. After final parameter adjustments, the test set is used to evaluate the performance of the model comprehensively, assessing for issues such as overfitting or underfitting. If no issues are identified, the model can be applied to the project.

Common supervised algorithms include partial least squares discriminant analysis (PLS-DA), decision trees (DTs), random forest (RF), support vector machine (SVM), K-Nearest Neighbor (kN), and eXtreme Gradient Boosting (XGBoost).

3.1.2.1 partial least squares discriminant analysis (PLS-DA)

Partial least squares discriminant analysis (PLS-DA) is also a dimensionality reduction algorithm. Unlike PCA, PLS is a "supervised" mode of partial least squares analysis, meaning that the grouping relationships of the samples are known, allowing for better selection of characteristic variables that distinguish each group and determination of relationships between the samples. DA stands for discriminant analysis. PLS-DA employs the partial least squares regression method to "reduce the dimensionality" of the data, establish a regression model, and conduct discriminant analysis on the regression results. PLS-DA is particularly suitable for selecting and interpreting metabolite signatures when studying biological systems (D'Andrea et al., 2023).

In an MS-based urine proteomics study of gastric lesions by Fan et al. (2022), PLS-DA was used to analyze the proteomics data of different groups and screen out 139 differential proteins with VIP>1.

In the study by D'Andrea et al. (2023), the PLS-DA model was confirmed through cross-validation, and the average variable importance in projection (VIP) score was used to identify metabolites that differed among sample classes.

3.1.2.2 Decision trees (DTs)

Decision trees (DTs) employ ML techniques to address classification and prediction problems. Nodes and leaves are the primary elements that form a decision tree (Chaubey et al., 2020). Nodes test specific properties, and leaves represent a class (Mesarić and Šebalj, 2016). Common decision trees include the ID3 tree, the C4.5 tree (information gain rate), and the CART tree (Gini coefficient) (Ross, 1993; Leo et al., 1984).

The ID3 algorithm is one of the classic decision tree algorithms. The C4.5 algorithm is an improvement upon the ID3 algorithm and can handle discontinuous features (Navada et al., 2011). The ID3 and the C4.5 algorithm are primarily used to address classification problems, but cannot be used to apply regression problems (Singh and Giri, 2014). The CART algorithm can manage both classification and regression problems.

Fannes et al. (2013) introduced CP-DT (Decision Tree Cleavage Prediction), an algorithm based on an ensemble of decision trees trained on publicly available peptide identification data from the PRIDE database. The study demonstrated that CP-DT can accurately predict trypsin cleavage (Fannes et al., 2013).

Decision tree algorithms are fast, however, they are generally not as accurate as other models.

3.1.2.3 Random forest (RF)

Random forest (RF) is a regression tree technique that employs bootstrap aggregation and predictor randomization to achieve a degree of predictive accuracy (Steven, 2017). Proposed by Breiman in 2001, RF employs randomization to create numerous decision trees and is a widely used tool for classification and regression in bioinformatics and related fields (Steven, 2017; Janitza et al., 2016). Compared to a single decision tree, a random forest exhibits stronger generalization performance. In classification problems, the outputs of these decision trees are voted and aggregated into one output; in regression problems, they are averaged and aggregated into one output (Steven, 2017). RF classification is a widely used supervised learning method for developing predictive models in many research settings (Speiser et al., 2019).

The random forest algorithm is simple and easy to implement, applicable to both classification and regression problems (Steven, 2017). It has the following features.

a. It can handle numerous input variables, and the more data features present, the more stable the model (Belgiu and Drăguț, 2016).
b. It can evaluate feature importance while determining the category (Archer and Kimes, 2008; Khalilia et al., 2011).
c. It can estimate valuable data and maintain a certain degree of accuracy even when a significant portion of the data is missing.

Khalilia et al. (2011) utilized National Inpatient Sample (NIS) data from the Healthcare Cost and Utilization Project (HCUP) to train RF classifiers for predicting eight disease categories. The results demonstrated good performance (Khalilia et al., 2011).

However, RF does not perform as well for regression problems as it does for classification and may not produce good classification results for small or low-dimensional datasets (datasets with fewer features).

3.1.2.4 Support vector machine (SVM)

Support vector machine (SVM) is a supervised algorithm that learns from examples to assign labels to objects (Boser et al., 1992). Compared to other ML methods, SVM is highly effective at identifying subtle patterns in complex data sets (Aruna and SP, 2011). The purpose of SVM is to create a decision boundary between two categories, facilitating the prediction of a label based on one or more feature vectors (WS, 2006). This decision boundary, called a hyperplane, should be oriented as far away as possible from the nearest data point for each class, referred to a support vector (Huang et al., 2018).

The computational complexity of SVM depends on the number of support vectors rather than the dimension of the sample space, thereby avoiding the "curse of dimensionality" (Markowetz, 2001). However, SVM is sensitive to missing data and is difficult for solving multi-classification problems (Cervantes et al., 2020). In areas where SVM performs poorly, researchers have developed other applications such as SVM for large datasets, multiple classifications, and imbalanced datasets (Cervantes et al., 2020).

Mavrogeorgis et al. (2023) obtained urine peptide data of 1850 healthy controls (HC) and CKD (diabetic nephropathy-DKD, IgA nephropathy-IgAN, vasculitis) participants from the Human Urine Proteome Database. UMAP was combined with SVM for binary (DKD, HC) and multi-class (DKD, HC, IgAN, vasculitis) classification.

3.1.2.5 K-Nearest neighbor (k-NN)

K-Nearest Neighbor (k-NN) is a simple and practical supervised learning algorithm frequently used to deal with classification problems (Boateng et al., 2020). It examines the k nearest sample points closest to the new sample point in the training set, using a specific distance metric, and classifies the new sample point into the category with the most occurrences among the k sample points (Abu Alfeilat et al., 2019). The parameter k is crucial, and its value should be optimally chosen (Zhang et al., 2017a). A value that is too low will increase the error rate, while a value that is too high can render the model ineffective (Zhang et al., 2017b).

The algorithm is simple in principle, easy to understand and implement, applicable to multi-classification problems, and requires no additional processing (Chaubey et al., 2020). However, k-NN involves substantial computational effort and requires considerable memory resources. Its performance is influenced by the parameter k and it tends to perform poorly on unbalanced datasets.

3.1.2.6 eXtreme gradient boosting (XGBoost)

eXtreme Gradient Boosting (XGBoost) is a machine learning model built on a decision tree ensemble and is among the most widely used machine learning algorithms (Kavzoglu and Teke, 2022). The algorithm has the following features:

a. It excels in processing both structured and unstructured data, frequently achieving higher accuracy compared to other algorithms (Arif Ali et al., 2023).
b. It relies on on decision tree integration, offers excellent interpretability, and provides insights into the importance of each feature (Kavzoglu and Teke, 2022).
c. It employs parallel computing technology and demonstrates high computational efficiency in processing large-scale data (Nalluri et al., 2020).

Li et al. (2024) developed a model incorporating 17 feature variables using XGBoost, based on the multidimensional data from a retrospective cohort of 274 papillary thyroid carcinoma (PTC) patients. This model demonstrated strong predictive performance in differentiating between low-risk and medium/high-risk PTC cases and was designated as the PTC Preoperative Risk Assessment Classifier (PRAC-PTC).

However, it is sensitive to parameters settings, with the choice of parameters significantly influencing the results (Demir and Şahin, 2022). In some cases, XGBoost may be overly complex and prone to overfitting the training data.

3.2 Applications of machine learning in MS-based multi-omics in thyroid disease

3.2.1 Applications in MS-based multi-omics data analysis

Extracting valuable insights from MS-based multi-omics data presents a significant challenge in bioinformatics (Tang et al., 2019). The complexity and high dimensionality of MS-based multi-omics datasets make traditional analysis methods challenging (Krassowski et al., 2020). Combining ML methods with MS-based multi-omics analysis mainly involves integrating various ML techniques to manage the complexity and volume of multi-omics data, aiming to enhance both accuracy and interpretability.

3.2.1.1 Missing data imputation

Missing data refers to the situation where data is incomplete due to some reasons during the process of data collection, transmission, and processing (Du et al., 2020). It is a common problem in MS-based omics data analysis (Huang et al., 2023). The simplest way to deal with missing values is to remove samples with missing values. However, if there are many missing values, such as the missing data of LC-MS-based omics data may be in the range of 30%–50%, a large number of samples will be eliminated, resulting in the loss of more useful information (Liebal et al., 2020).

Imputation methods provide an alternative way of handling missing data rather than discarding missing values and associated

data (Huang et al., 2023). The mean, median, mode, etc., of the feature can be used to fill the missing values (Emmanuel et al., 2021). However, these simple methods do not consider the relationship between data variables, which sometimes makes the results of data analysis unreliable. Among the methods for dealing with missing values, many other filling methods consider the relationship between data variables (Baraldi and Enders, 2010). ML algorithms, such as regression, k-NN, and RF, can help resolve missing data problems in multi-omics datasets by inferring values based on observed patterns in existing data (Emmanuel et al., 2021; Mirza et al., 2019).

3.2.1.2 Dimensionality reduction

MS-based multi-omics data may have multiple layers of variables and a large number of attributes, so-called high-dimensional data (Arjmand et al., 2022). While high-dimensional data will cause great trouble for subsequent data processing (Cao and Lin, 2015), dimensionality reduction is a crucial step (Fanaee and Thoresen, 2019). It aims to reduce the number of variables considered, making the data more manageable and easier to analyze while retaining as much information as possible. Before applying dimensionality reduction, multi-omics data need to be preprocessed, such as normalization and missing value filling, to ensure that the data is in a form suitable for analysis (Reska et al., 2021). Dimensionality reduction improves computational efficiency, reduces noise while retaining important information, facilitating data processing (Alhassan and Wan Zainon, 2021). Many ML algorithms can facilitate data processing by reducing data dimensionality while retaining important information, such as PCA, t-SNE, PLS-DA, and UMAP.

3.2.1.3 Clustering and classification

Clustering is an unsupervised learning method that groups data based on the attributes of the input features (Reel et al., 2021). Classification is a supervised learning method that provides predicted output as a discrete class (Reel et al., 2021). ML algorithms can group samples into clusters or classify them based on distinct patterns present in multi-omics data which can discover subtypes or stratify patients and identify similarities among clustered patients (Goecks et al., 2020).

3.2.1.4 Feature selection

Feature selection is a key step in multi-omics analysis and helps reduce data dimensionality. In this sense, feature selection has similar motivations to dimensionality reduction as described above. Feature selection can remove irrelevant features and reduce the number of features used in the analysis, thereby reducing the difficulty of the learning task (Li et al., 2017a). It should be noted that the feature selection process must ensure that no important features are lost.

For a multi-omics dataset, a set of attributes is included, some of them may be critical and useful, while others may be useless. Attributes are called features, those that are useful for the current learning task are called relevant features, and those that are useless are called irrelevant features (Kotsiantis, 2011). The process of selecting a subset of relevant features from a given set of features is called feature selection (Jović et al., 2015).

Feature selection is an important data preprocessing process. In real machine learning tasks, feature selection is usually performed after obtaining multi-omics data, and then the learner is trained.

In practical applications, feature selection methods are mainly divided into filter, wrapper, and embedded methods (Venkatesh and Anuradha, 2019).

a. Filter Selection

Filter selection first selects features from the data set and then trains the learner (Li et al., 2017a). The feature selection process is independent of the subsequent learner. It is the simplest and most commonly used method to implement feature selection. The core of the filtering method selection is to sort the features according to their value, to achieve the selection or elimination of any proportion/quantity of features.

Filter selection is computationally efficient and relatively simple to implement, but it ignores the interactions between features (Jiliang et al., 2014).

b. Wrapper Selection

Unlike filter-based feature selection, which does not consider subsequent learners, wrapper selection directly uses the performance of the most important learner as the evaluation criterion for the feature subset (Wald et al., 2013; Cadenas et al., 2013).

Since wrapper selection directly optimizes a given learner, it is better than filter selection in terms of the final learner performance (Cadenas et al., 2013). However, since the learner needs to be trained multiple times during the feature selection process, it requires a huge amount of computation.

c. Embedded selection

Embedded selection integrates the feature selection process with the learner training process, that is, feature selection is automatically performed during the learner training process (Li et al., 2017b). The most commonly used are tree models and a series of ensemble algorithms based on tree models because the model provides important information about feature importance.

Embedded selection combines the advantages of filter selection and wrapper selection in terms of computational cost and performance, but it cannot identify highly relevant features (Remeseiro and Bolon-Canedo, 2019).

3.2.2 Applications in clinical researches

The contribution of ML to thyroid disease is not only in processing data, but also in the classification, clinical diagnosis, treatment, prognosis and risk stratification of thyroid diseases. In clinical researches, ML is beneficial and essential (Komuro et al., 2023). ML enables efficient analysis of extensive data sets and improve disease diagnosis and classification by building predictive models of disease, and has a particularly important impact on improving MS-based clinical multi-omics research (Perakakis et al., 2018; Arjmand et al., 2022). ML models primarily select important features from complex data to construct predictive models and output data as predictive labels based on identified patterns (Ngan et al., 2023). Using models like decision trees, RF, SVM, and XGBoost to predict outcomes based on data facilitates predictive models, including disease

diagnosis, prognosis, and treatment, based on multi-omics data (Quazi, 2022; Wekesa and Kimwele, 2023). Validation and evaluation are critical steps in ML for multi-omics analysis to ensure that the models and methods used are reliable and robust. There are several approaches:

Cross-Validation: Use k-fold cross-validation to assess model performance.

External Validation: Validate results with independent datasets to ensure robustness.

Performance Metrics: Use appropriate metrics to evaluate the model (e.g., accuracy, precision, recall, area under curve (AUC), F1-score for classification; R 2, RMSE for regression).

The trained and validated ML models are expected to be able to assess cancer risk and facilitate the development of preoperative cancer diagnosis (Ngan et al., 2023).

3.2.2.1 Diagnosis and classification

The classification model is the most commonly used type of ML models in clinical researches which can output the type of thyroid disease it predicts based on the input features, such as clinical factors and omics data, thereby providing insights into clinical diagnosis (Figure 4). For thyroid disease prediction, ML methods have been applied in various existing research works. Prediction of thyroid disease at its early stages and categorization into cancer or other thyroid disease is very helpful for treating and recovering the maximum number of patients (Gupta et al., 2024). MS-based multi-omics have become a powerful technique for biomarker discovery which is significant for clinic (Torun et al., 2022). However, screening biomarkers from complex MS data requires reliable bioinformatics tools and ML can be a good choice (Torun et al., 2022). Analyzing large and complex datasets by ML models enables the identification of subtle biomarkers and disease signatures, which leads to earlier and more accurate diagnoses (Ng et al., 2023). This is particularly beneficial in complex diseases specially cancer, where early detection can significantly improve prognosis. The integration of ML and MS-based multi-omics not only improves the efficiency of biomarker discovery but also helps develop more accurate and reliable diagnostic tools, ultimately advancing the field of precision medicine and improving patient outcomes (Johnson et al., 2021; Krishnan et al., 2023).

From the perspective of classification, existing researches mainly focused on binary classification problems in thyroid disease classification based on ML models. Kumari et al. (2024) used age, gender, and hormone levels as features and combined them with 3 ML models to classify hyperthyroidism and hypothyroidism. Results showed that XGBoost was the top-performing model for this task (Kumari et al., 2024).

Al-muwaffaq and Bozkus (2016) developed the Machine Learning tool for Thyroid Disease Diagnosis (MLTDD) mainly focused on thyroid gland medical diseases caused by underactive or overactive thyroid glands. The prediction accuracy was in range between 98.7% and 99.8% for testing. MLTDD can effectively help to make the right clinical decision.

Xi et al. (2022) constructed 6 ML models to predict the malignancy of thyroid nodules. RF and Gradient Boosting Machine (GBM) showed better overall diagnostic accuracy and ability to identify malignant nodules. Their method can be used as additional evidence in the preoperative diagnosis of thyroid cancer.

Guo et al. (2022b) constructed a diagnostic model of benign and malignant thyroid tumors. The benign group contained 5 thyroid diseases and the malignant group contained six thyroid diseases. Clinical factors were used as features and RF, XGBoost, LightGBM, and AdaBoost models were constructed. RF model showed the best performance. Their research proposed a model incorporating novel biomarkers which could be a powerful and promising tool for predicting benign and malignant thyroid tumors (Guo et al., 2022b).

Recently, some researches have begun to focus on multi-classes classification problems. Chaganti et al. (2022) used ML models to predict five thyroid diseases. Results showed that the extra tree classifier-based selected feature yields the best results with 0.99 accuracy and an F1 score when used with the RF classifier (Chaganti et al., 2022).

Gupta et al. (2024) used ML models to predict ten thyroid diseases. Results suggested that an accuracy of 0.998 can be obtained using the optimized AdaBoost model by differential evolution (Gupta et al., 2024). These researches applied ML models to a variety of thyroid diseases, which helped to diagnose diseases more conveniently and comprehensively in clinical practice.

ML models can not only be used to classify the types of thyroid disease, but also to determine whether there is a certain gene mutation in the disease. Kwon et al. (2020) used ML models combined with radiomics to predict the presence or absence of B-Raf proto-oncogene, serine/threonine kinase (BRAF) mutation in PTC and the results showed that the classification accuracy of these models was higher than 60%. Although this research provided a new perspective of application of ML in thyroid disease which focused on the gene mutation, the classification model did not show excellent performance in predicting the presence of BRAF mutation in PTC and need to be further validated in a larger dataset to better assess their potential clinical use.

In addition, ML models can also be used to predict treatment trends and prognosis. Aversano et al. (2021) combined ten different ML models with parameters related to the person being treated to predict whether the LT4-based treatment needs to be increased or decreased for the patients with hypothyroidism. This study provides reference insights for clinical treatment plans.

Wang et al. (2024a) presented ML models based on comprehensive predictors to predict the structural recurrence risk of PTC patients. All the patients were treated with thyroid surgery and radioiodine. Twenty-nine perioperative variables consisting of four dimensions (demographic characteristics and comorbidities, tumor-related variables, lymph node-related variables, and metabolic and inflammatory markers) were analyzed (Wang et al., 2024a). The results showed that the RF model achieved the expected prediction effect, with good discrimination, calibration, and interpretability, and revealed the potential of ML models in improving the accuracy of risk stratification of PTC patients (Wang et al., 2024a).

It is also an important task for ML models to classify clinical patient samples based on MS-based multi-omics data. In many cases, it can be used to determine which ion m/z values contribute most to the models, thereby facilitating the identification of biomarkers and further facilitating the diagnosis and classification of thyroid disease (Beck et al., 2024). Recently, an increasing number of researches have focused on combining ML with MS data for thyroid disease classification, even for further treatment strategies, which have provided broader prospect.

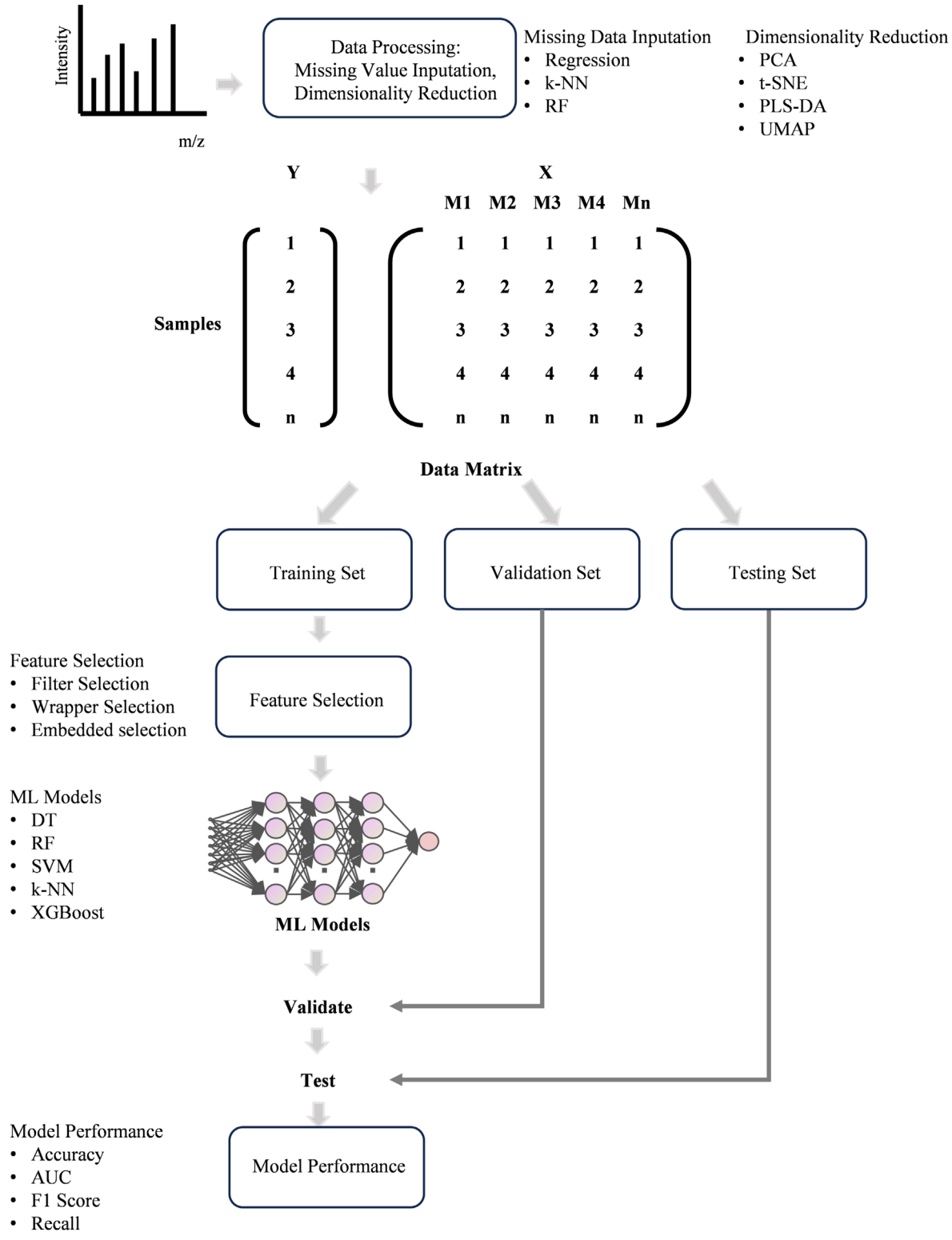

FIGURE 4
The workflow of ML for data processing and modeling.

Sun et al. (2022) proposed the first protein-based neural network classifier for thyroid nodules. Their research was the first to establish a deep proteome data repository for various thyroid lesions, analyzed a larger sample size and obtained deeper proteome coverage (Sun et al., 2022). The large-scale thyroid proteome map combined with the neural network model demonstrated the power

of the classifier and is expected to be quickly applied to clinical practice to supplement the deficiencies of traditional cytopathology (Sun et al., 2022). Their recent work (Sun et al., 2024) established a protein-based model with targeted MS for the diagnosis of FTC and FTA. This model used 24 proteins filtered by XGBoost as features and performed better than gene-based model. The protein model has 95.7% negative predictive value for ruling out malignant nodules (Sun et al., 2024).

Wang et al. (2022b) develop a rapid classification method by ML and MS-based metabolomics to diagnose PTC. The metabolomics of frozen samples were performed by probe electrospray ionization (PESI) mass and SVM and RF models were used. For the classification of PTC from PTC adjacent tissues, SVM performed better than RF. Their another work developed a rapid method to classify the malignant and benign thyroid nodules by PESI-MS-based machine learning (Wang et al., 2022c). For each FNAB sample, only 10 min is needed to determine its malignancy, which is much easier and faster than traditional diagnosis (Wang et al., 2022c).

Chen et al. (2024) explored the potential of rapid thyroid disease screening using the ZrMOF/Au-assisted LDI-MS platform, enabling rapid screening of malignant thyroid disease from benign patients. The authors constructed a panel of 43 key metabolites as features for ML models to discriminate thyroid cancer from thyroid nodules and NN, RF, LR, and SVM models were used for classification (Chen et al., 2024). The results showed that NN had the best classification performance.

Zhu et al. (2024) combined ML with MS-based proteomics and selected four proteins as features with the highest contributions to predict the efficacy of iodine therapy. This research showed the ability to pre-identify PTC patients who are resistant to radioactive iodine therapy (Zhu et al., 2024).

The above articles were retrieved by entering "machine learning" and "thyroid disease" in Google Scholar and some representative ones involving "application of ML in diagnosis of thyroid disease" were selected by us. We showed them in Table 1 in the order we mentioned in this section with three summary measures, sample types, features, and supervised ML models.

By analyzing the above research works, we found that patients' personal data and hormone parameters are often used as features, such as age, sex, thyroid-stimulating hormone (TSH), total serum triiodothyronine (T3), thyroid binding globulin (TBG) and total serum thyroxin (T4). Among these features, almost every researcher has selected some features for thyroid disease diagnosis work. MS-based proteomics and metabolomics data are also often used but almost only single type is used. In order to analyze thyroid disease more comprehensively, combining metabolomics and proteomics or combining them with other omics data such as genomics and clinical indicators should be further considered in future studies.

3.2.2.2 Risk stratification

A significant advantage of ML in the clinic is its ability to facilitate early intervention (Adlung et al., 2021). By using advanced algorithms and predictive models, ML algorithms can detect potential health risks at an early stage, and effectively classify diseases into different risk categories, which offer a nuanced understanding of the likelihood of patients developing specific health conditions and allowing clinicians to intervene promptly and implement targeted treatments (Beaulieu-Jones et al., 2021; Choudhury and Asan, 2020; Sun et al., 2023). This goes beyond traditional diagnostic methods, providing a more personalized and proactive approach to healthcare (Yadav, 2024).

Saima Sharleen Islam (Sun et al., 2023) et al. used 11 ML models to predict thyroid risk and used accuracy and recall as evaluation indicators. The results show that the ANN classifier outperforms the others in terms of accuracy. Zhao et al. (2022a) et al. used 5 ML models to predict the risk of nodular thyroid disease in coal miners, with the XGB model having the best overall predictive performance.

Wang et al. (2022a) presented ML models based on comprehensive predictors to predict the structural recurrence risk of PTC patients. All the patients were treated with thyroid surgery and radioiodine. Twenty-nine perioperative variables consisting of four dimensions (demographic characteristics and comorbidities, tumor-related variables, lymph node-related variables, and metabolic and inflammatory markers) were analyzed Wang et al. (2022a). The results showed that the RF model achieved the expected prediction effect, with good discrimination, calibration, and interpretability, and revealed the potential of ML models in improving the accuracy of risk stratification of PTC patients Wang et al. (2022a).

MS-based multi-omics data can also be combined with ML for risk stratification in thyroid disease. Li et al. (2024) first reported a Preoperative Risk Assessment Classifier for PTC (PRAC-PTC) which constructed by ML models used clinical indicators, immune indices, genetic feature, and MS-based proteomics as multidimensional features. The results showed that six proteins (DPP7, PDLIM3, Col12A1, CTSL, TUBB2A, and ITGB5) were identified as the best discriminable proteins between low-risk and intermediate-risk/high-risk PTCs (Li et al., 2024). XGBoost showed the best performance among these ML models authors used. PRAC-PTC can increase the accuracy of the preoperative risk stratification and decrease unnecessary surgery or overtreatment.

Wang et al. (2024b) used different ML algorithms to explore the relationship between mixed-semi-volatile organic compounds (SVOCs) exposure and thyroid nodule. The data was collected by GC-MS/MS. RF and AdaBoost models were selected to screen out the features based on their contribution to the models. Weighted quantile sum (WQS) regression and Bayesian kernel machine regression (BKMR) were used to assess the mixed effects of the SVOCs exposure on thyroid nodule (Wang et al., 2024b). The results showed that high levels of exposure to SVOCs increase the risk of PTC and nodular goiters (NG), with Fluazifop-butyl and Fenpropathrin playing a major role (Wang et al., 2024b).

Wang et al. (2024c) proposed a ML-based objective method to individual to predict the risk of pediatric papillary thyroid carcinomas (PPTCs). They collected the clinical factors and MS-based proteomics data and nineteen proteins were selected by ML models to construct a protein-based personalized prognostic prediction model which can stratify PPTC patients into high- or low-recurrence risk groups and provide a suggestion for clinical decision-making and individualized treatment.

The combination of ML and MS-based multi-omics data in thyroid disease not only enhances disease risk assessment but also improves the approach to patient care. Through early intervention, patient stratification, and the implementation of targeted preventive measures, ML makes a significant contribution to improving patient health and building a more personalized and efficient clinical treatment system (Khalifa and Albadawy, 2024; Khalifa et al., 2024).

TABLE 1 Applications of machine learning in diagnosis of thyroid disease.

References	Year	Sample types	Features	Supervised ML models
Gupta et al. (2024)	2024	Concurrent Non-thyroid Illness, Compensated Hypothyroid, Increased Binding Protein, Primary Hypothyroid, etc	not given	RF, SVM, LR, ADA, GBM, CNN, RNN, LSTM
Kumari et al. (2024)	2024	Hyperthyroidism and Hypothyroidism	age, sex, pregnancy, T3, T4, and TSH	RF, SVM, XGB
Al-muwaffaq and Bozkus (2016)	2016	Hyperthyroidism, Hypothyroidism and normal function of the thyroid gland	age, sex, on_thyroxine, query_on_thyroxine, on_antithyroid_medication, sick, pregnant, TSH, T3, T4, T4U, FTIetc.	DT
Xi et al. (2022)	2022	Thyroid Cancer	demographic information, ultrasound features, and blood test results	RF, SVM, LR, LDA, GBM
Guo et al. (2022b)	2022	Benign and Malignant Thyroid Tumors	Peripheral blood indicators, BRAFV600E gene, demographic indicators	RF, XGB, GBM, ADA
Chaganti et al. (2022)	2022	Hashimoto's thyroiditis, binding protein (increased binding protein), Autoimmune Thyroiditis, and Non-Thyroidal Syndrome (NTIS)	age, sex, on_thyroxine, query_on_thyroxine, on_antithyroid_medication, sick, pregnant, TSH, T3, T4, T4U, FTI, TBGetc.	RF, LR, SVM, ADA, GBM
Kwon et al. (2020)	2020	PTC	radiomics features	LR, RF, SVM
Aversano et al. (2021)	2021	Hypothyroidism	personal information, family history, physical characteristics, hormonal and thyroid parameters, parameters relating to blood tests	ADA, XGB, GBM, CAT, DT, RF, ExtraTree, k-NN, Naive Bayes, MLP
Wang et al. (2024a)	2024	PTC	demographic characteristics and comorbidities, tumor-related variables, lymph node (LN)-related variables, metabolic and inflammatory markers	LR, XGB, RF, SVM, NN
Sun et al. (2022)	2022	Thyroid Nodules	MS-based proteomics	NN
Sun et al. (2024)	2024	FTA and FTC	MS-based proteomics	XGB
Wang et al. (2022b)	2022	PTC	MS-based metabolomics	RF, SVM
Wang et al. (2022c)	2022	Thyroid Nodules	MS-based ions	RF, SVM, MLP
Chen et al. (2024)	2024	Thyroid Cancer and Thyroid Nodules	MS-based metabolomics	NN, RF, LR, SVM
Zhu et al. (2024)	2024	Radioactive Iodine Refractory (RAIR) and Non-Radioactive Iodine Refractory (Non-RAIR) PTC	MS-based proteomics	XGB

ADA, adaptive boosting; CAT, CatBoosting; CNN, convolutional neural network; DT, decision tree; GBM, gradient boosting machine; k-NN, k-nearest neighbor; LDA, linear discriminant analysis; LR, logistic regression; LSTM, long short-term memory; MLP, multilayer perceptron; NN, neural network; RF, random forest; RNN, recurrent neural network; SVM, support vector machine; XGB, eXtreme Gradient Boosting.

4 Summary

In the 21st century, the development of medical science has entered the era of big data, with ML algorithms, as a cornerstone of artificial intelligence, beginning to emerge in clinical disease research (Li et al., 2023b). From the clinical perspective, the application of ML in thyroid disease can contribute to the classification, diagnosis, treatment and prognosis. MS-based multi-omics analysis utilizing ML technology, offers significant prospects for early disease detection and prevention. However, there are currently limited examples of successful application of such technologies into clinical practice (Kelly et al., 2019). There are also many challenges must be addressed. In the biomarker field, the primary challenge is not in data analysis, but in the

collection of comprehensive clinical data for each particular patient (Desaire et al., 2022; Arzyeh et al., 2020). The success of ML algorithms in the medical field largely depends on the quality, diversity, and completeness of the training data (Sarker, 2021). While the number of samples in some researches are not small, collecting more data will increase the diversity of patients. Models trained on broader and more diverse datasets will generalize better to new patients when deployed in real-world scenarios (Xi et al., 2022). Most researches mainly use publicly available datasets, and many datasets have the problem of class imbalance, with a very small number of samples in a certain class. Meanwhile, this phenomenon also leads to the results not being universal. When ML models are applied to such datasets, the models will overfit to the majority class, resulting in incorrect predictions for the minority class. The other limitation is that only a few types of thyroid diseases have been used to classification problems in existing studies, and most studies focus on binary classification, which has caused certain limitations in the application of ML to clinical applications of thyroid diseases. In disease research, this entails collecting detailed information about a patient's medical history, genetic profile, treatment response, and long-term outcomes. Furthermore, patient privacy and data security must be carefully managed and addressed. As data volume increases, it is imperative to implement robust measures to protect patient privacy and uphold ethical standards (Abouelmehdi et al., 2018).

In summary, while ML holds substantial for clinical disease researches, it also encounters significant challenges to widespread adoption. Addressing these challenges will pave the way for realizing the full potential of ML in enhancing disease diagnosis, prognosis, and personalized treatment strategies.

Author contributions

YC: Writing–original draft, Writing–review and editing. MZ: Writing–review and editing. YG: Writing–review and editing. ZZ: Writing–review and editing, Resources. XZ: Conceptualization, Writing–review and editing, Supervision, Funding acquisition, Resources.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

Thanks are given to ZZ for assistance with histopathologic information for endocrine tumors and to XZ for valuable discussions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abouelmehdi, K., Beni-Hessane, A., and Khaloufi, H. (2018). Big healthcare data: preserving security and privacy. *J. Big Data* 5 (1), 1. doi:10.1186/s40537-017-0110-7

Abu Alfeilat, H. A., Hassanat, A. B. A., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M. B., Eyal Salman, H. S., et al. (2019). Effects of distance measure choice on K-nearest neighbor classifier performance: a review. *Big Data* 7 (4), 221–248. doi:10.1089/big.2018.0175

Adlung, L., Cohen, Y., Mor, U., and Elinav, E. (2021). Machine learning in clinical decision making. *Med* 2 (6), 642–665. doi:10.1016/j.medj.2021.04.006

Alber, M., Buganza Tepole, A., Cannon, W. R., De, S., Dura-Bernal, S., Garikipati, K., et al. (2019). Integrating machine learning and multiscale modeling-perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. *NPJ Digit. Med.* 2, 115. doi:10.1038/s41746-019-0193-y

Alhassan, A. M., and Wan Zainon, W. M. N. (2021). Review of feature selection, dimensionality reduction and classification for chronic disease diagnosis. *IEEE Access* 9, 87310–87317. doi:10.1109/access.2021.3088613

Al-muwaffaq, I., and Bozkus, Z. (2016). MLTDD: use of machine learning techniques for diagnosis of thyroid gland disorder. *Comput. Sci. and Inf. Technol. (CS and IT)*, 67–73. doi:10.5121/csit.2016.60507

Alseekh, S., Aharoni, A., Brotman, Y., Contrepois, K., D'Auria, J., Ewald, J., et al. (2021). Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. *Nat. Methods* 18 (7), 747–756. doi:10.1038/s41592-021-01197-1

Angra, S., and Ahuja, S. (2017). "Machine learning and its applications: a review," in 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), 23-25 March 2017, 57–60. doi:10.1109/icbdaci.2017.8070809

Archer, K. J., and Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Comput. Statistics and Data Analysis* 52 (4), 2249–2260. doi:10.1016/j.csda.2007.08.015

Arif Ali, Z. H., Abduljabbar, Z., Tahir, H., Bibo Sallow, A., and Almufti, S. M. (2023). eXtreme gradient boosting algorithm with machine learning: a review. *Acad. J. Nawroz Univ.* 12 (2), 320–334. doi:10.25007/ajnu.v12n2a1612

Arjmand, B., Hamidpour, S. K., Tayanloo-Beik, A., Goodarzi, P., Aghayan, H. R., Adibi, H., et al. (2022). Machine learning: a new prospect in multi-omics data analysis of cancer. *Front. Genet.* 13, 824451. doi:10.3389/fgene.2022.824451

Aruna, S., and Sp, R. (2011). A novel SVM based CSSFFS feature selection algorithm for detecting breast cancer. *Int. J. Comput. Appl.* 31 (8), 14–20. doi:10.5120/3844-5346

Auslander, N., Gussow, A. B., and Koonin, E. V. (2021). Incorporating machine learning into established bioinformatics frameworks. *Int. J. Mol. Sci.* 22 (6), 2903. doi:10.3390/ijms22062903

Aversano, L., Bernardi, M. L., Cimitile, M., Iammarino, M., Macchia, P. E., Nettore, I. C., et al. (2021). Thyroid Disease Treatment prediction with machine learning approaches. *Procedia Comput. Sci.* 19, 1031–1040. doi:10.1016/j.procs.2021.08.106

Babu, M., and Snyder, M. (2023). Multi-omics profiling for health. *Mol. Cell. Proteomics* 22 (6), 100561. doi:10.1016/j.mcpro.2023.100561

Ball, L. E., Agana, B. A., Comte-Walters, S., Bethard, J. R., and Burnette, B. B. (2023). An introduction to mass spectrometry-based proteomics. *Encycl. Cell. Biol.*, 132–140. doi:10.1016/b978-0-12-821618-7.00143-7

Baraldi, A. N., and Enders, C. K. (2010). An introduction to modern missing data analyses. *J. Sch. Psychol.* 48 (1), 5–37. doi:10.1016/j.jsp.2009.10.001

Basolo, F., Macerola, E., Poma, A. M., and Torregrossa, L. (2023). The 5(th) edition of WHO classification of tumors of endocrine organs: changes in the diagnosis of follicular-derived thyroid carcinoma. *Endocrine* 80 (3), 470–476. doi:10.1007/s12020-023-03336-4

Beattie, J. R., and Esmonde-White, F. W. L. (2021). Exploration of principal component analysis: deriving principal component analysis visually using spectra. *Appl. Spectrosc.* 75 (4), 361–375. doi:10.1177/0003702820987847

Beaulieu-Jones, B. K., Yuan, W., Brat, G. A., Beam, A. L., Weber, G., Ruffin, M., et al. (2021). Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? *NPJ Digit. Med.* 4 (1), 62. doi:10.1038/s41746-021-00426-3

Becht, E., McInnes, L., Healy, J., Dutertre, C. A., Kwok, I. W. H., Ng, L. G., et al. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* 37, 38–44. doi:10.1038/nbt.4314

Beck, A. G., Muhoberac, M., Randolph, C. E., Beveridge, C. H., Wijewardhane, P. R., Kenttamaa, H. I., et al. (2024). Recent developments in machine learning for mass spectrometry. *ACS Meas. Sci. Au* 4 (3), 233–246. doi:10.1021/acsmeasuresciau.3c00060

Belgiu, M., and Drăguţ, L. (2016). Random forest in remote sensing: a review of applications and future directions. *ISPRS J. Photogrammetry Remote Sens.* 114, 24–31. doi:10.1016/j.isprsjprs.2016.01.011

Berger, M. F., and Mardis, E. R. (2018). The emerging clinical relevance of genomics in cancer medicine. *Nat. Rev. Clin. Oncol.* 15 (6), 353–365. doi:10.1038/s41571-018-0002-6

Biomarkers, D. W. G. (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin. Pharmacol. Ther.* 69 (3), 89–95. doi:10.1067/mcp.2001.113989

Blekherman, G., Laubenbacher, R., Cortes, D. F., Mendes, P., Torti, F. M., Akman, S., et al. (2011). Bioinformatics tools for cancer metabolomics. *Metabolomics* 7 (3), 329–343. doi:10.1007/s11306-010-0270-3

Boateng, E. Y., Otoo, J., and Abaye, D. A. (2020). Basic tenets of classification algorithms K-Nearest-Neighbor, support vector machine, random forest and neural network: a review. *J. Data Analysis Inf. Process.* 08 (04), 341–357. doi:10.4236/jdaip.2020.84020

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proc. fifth Annu. workshop Comput. Learn. theory*, 144–152. doi:10.1145/130385.130401

Cadenas, J. M., Garrido, M. C., and Martínez, R. (2013). Feature subset selection Filter–Wrapper based on low quality data. *Expert Syst. Appl.* 40 (16), 6241–6252. doi:10.1016/j.eswa.2013.05.051

Califf, R. M. (2018). Biomarker definitions and their applications. *Exp. Biol. Med. (Maywood)* 243 (3), 213–221. doi:10.1177/1535370217750088

Cao, J., and Lin, Z. (2015). Extreme learning machines on high dimensional and large data applications: a survey. *Math. Problems Eng.* 2015, 1–13. doi:10.1155/2015/103796

Caria, P., Dettori, T., Frau, D. V., Lichtenzstejn, D., Pani, F., Vanni, R., et al. (2019). Characterizing the three-dimensional organization of telomeres in papillary thyroid carcinoma cells. *J. Cell. Physiol.* 234 (4), 5175–5185. doi:10.1002/jcp.27321

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., and Lopez, A. (2020). A comprehensive survey on support vector machine classification: applications, challenges and trends. *Neurocomputing* 408, 189–215. doi:10.1016/j.neucom.2019.10.118

Chaganti, R., Rustam, F., De La Torre Diez, I., Mazon, J. L. V., Rodriguez, C. L., and Ashraf, I. (2022). Thyroid disease prediction using selective features and machine learning techniques. *Cancers (Basel)* 14 (16), 3914. doi:10.3390/cancers14163914

Chatzimparmpas, A., Martins, R. M., and Kerren, A. (2020). t-viSNE: interactive assessment and interpretation of t-SNE projections. *IEEE Trans. Vis. Comput. Graph* 26 (8), 2696–2714. doi:10.1109/TVCG.2020.2986996

Chaubey, G., Bisen, D., Arjaria, S., and Yadav, V. (2020). Thyroid disease prediction using machine learning approaches. *Natl. Acad. Sci. Lett.* 44 (3), 233–238. doi:10.1007/s40009-020-00979-z

Chen, C., Wang, J., Pan, D., Wang, X., Xu, Y., Yan, J., et al. (2023b). Applications of multi-omics analysis in human diseases. *MedComm* 4 (4), e315. doi:10.1002/mco2.315

Chen, D. W., Lang, B. H. H., McLeod, D. S. A., Newbold, K., and Haymart, M. R. (2023a). Thyroid cancer. *Lancet* 401 (10387), 1531–1544. doi:10.1016/S0140-6736(23)00020-X

Chen, J., Yu, X., Qu, Y., Wang, X., Wang, Y., Jia, K., et al. (2024). High-performance metabolic profiling of high-risk thyroid nodules by ZrMOF hybrids. *ACS Nano* 18 (32), 21336–21346. doi:10.1021/acsnano.4c05700

Chen, X. W., and Gao, J. X. (2016). Big data bioinformatics. *Methods* 111, 1–2. doi:10.1016/j.ymeth.2016.11.017

Choudhury, A., and Asan, O. (2020). Role of artificial intelligence in patient safety outcomes: systematic literature review. *JMIR Med. Inf.* 8 (7), e18599. doi:10.2196/18599

Cieslak, M. C., Castelfranco, A. M., Roncalli, V., Lenz, P. H., and Hartline, D. K. (2020). t-Distributed Stochastic Neighbor Embedding (t-SNE): a tool for eco-physiological transcriptomic analysis. *Mar. Genomics* 51, 100723. doi:10.1016/j.margen.2019.100723

D'Andrea, G., Jing, L., Peyrottes, I., Guigonis, J. M., Graslin, F., Lindenthal, S., et al. (2023). Pilot study on the use of untargeted metabolomic fingerprinting of liquid-cytology fluids as a diagnostic tool of malignancy for thyroid nodules. *Metabolites* 13 (7), 782. doi:10.3390/metabo13070782

Da Silva Lopes, M. A., Doria Neto, A. D., and De Medeiros Martins, A. (2020). Parallel t-SNE applied to data visualization in smart cities. *IEEE Access* 8, 11482–11490. doi:10.1109/access.2020.2964413

Davis, K. D., Aghaeepour, N., Ahn, A. H., Angst, M. S., Borsook, D., Brenton, A., et al. (2020). Discovery and validation of biomarkers to aid the development of safe and effective pain therapeutics: challenges and opportunities. *Nat. Rev. Neurol.* 16 (7), 381–400. doi:10.1038/s41582-020-0362-2

Demir, S., and Şahin, E. K. (2022). Liquefaction prediction with robust machine learning algorithms (SVM, RF, and XGBoost) supported by genetic algorithm-based feature selection and parameter optimization from the perspective of data processing. *Environ. Earth Sci.* 81 (18), 459. doi:10.1007/s12665-022-10578-4

Desaire, H., Go, E. P., and Hua, D. (2022). Advances, obstacles, and opportunities for machine learning in proteomics. *Cell. Rep. Phys. Sci.* 3 (10), 101069. doi:10.1016/j.xcrp.2022.101069

Ding, J., and Feng, Y.-Q. (2023). Mass spectrometry-based metabolomics for clinical study: recent progresses and applications. *TrAC Trends Anal. Chem.* 158, 116896. doi:10.1016/j.trac.2022.116896

Du, J., Hu, M., and Zhang, W. (2020). Missing data problem in the monitoring system: a review. *IEEE Sensors J.* 20 (23), 13984–13998. doi:10.1109/jsen.2020.3009265

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., and Tabona, O. (2021). A survey on missing data in machine learning. *J. Big Data* 8 (1), 140. doi:10.1186/s40537-021-00516-9

Fan, H., Li, X., Li, Z. W., Zheng, N. R., Cao, L. H., Liu, Z. C., et al. (2022). Urine proteomic signatures predicting the progression from premalignancy to malignant gastric cancer. *EBioMedicine* 86, 104340. doi:10.1016/j.ebiom.2022.104340

Fanaee, T. H., and Thoresen, M. (2019). Multi-insight visualization of multi-omics data via ensemble dimension reduction and tensor factorization. *Bioinformatics* 35 (10), 1625–1633. doi:10.1093/bioinformatics/bty847

Fannes, T., Vandermarliere, E., Schietgat, L., Degroeve, S., Martens, L., and Ramon, J. (2013). Predicting tryptic cleavage from proteomics data using decision tree ensembles. *J. Proteome Res.* 12 (5), 2253–2259. doi:10.1021/pr4001114

Fernández, P. L., Merino, M. J., Gómez, M., Campo, E., Medina, T., Castronovo, V., et al. (1997). Galectin-3 and laminin expression in neoplastic and non-neoplastic thyroid tissue. *J. Pathol.* 181 (1), 80–86. doi:10.1002/(SICI)1096-9896(199701)181:1<80::AID-PATH699>3.0.CO;2-E

Galal, A., Talal, M., and Moustafa, A. (2022). Applications of machine learning in metabolomics: disease modeling and classification. *Front. Genet.* 13, 1017340. doi:10.3389/fgene.2022.1017340

Gild, M. L., Bullock, M., Pon, C. K., Robinson, B. G., and Clifton-Bligh, R. J. (2016). Destabilizing RET in targeted treatment of thyroid cancers. *Endocr. Connect.* 5 (1), 10–19. doi:10.1530/EC-15-0098

Gisbrecht, A., Schulz, A., and Hammer, B. (2015). Parametric nonlinear dimensionality reduction using kernel t-SNE. *Neurocomputing* 147, 71–82. doi:10.1016/j.neucom.2013.11.045

Goecks, J., Jalili, V., Heiser, L. M., and Gray, J. W. (2020). How machine learning will transform biomedicine. *Cell.* 181 (1), 92–101. doi:10.1016/j.cell.2020.03.022

Granato, D., Santos, J. S., Escher, G. B., Ferreira, B. L., and Maggio, R. M. (2018). Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: a critical perspective. *Trends Food Sci. and Technol.* 72, 83–90. doi:10.1016/j.tifs.2017.12.006

Guarino, V., Castellone, M. D., Avilla, E., and Melillo, R. M. (2010). Thyroid cancer and inflammation. *Mol. Cell. Endocrinol.* 321 (1), 94–102. doi:10.1016/j.mce.2009.10.003

Guo, J., Yu, H., Xing, S., and Huan, T. (2022a). Addressing big data challenges in mass spectrometry-based metabolomics. *Chem. Commun. (Camb)* 58 (72), 9979–9990. doi:10.1039/d2cc03598g

Guo, Y.-y., Li, Z.-j., Du, C., Gong, J., Liao, P., Zhang, J.-x., et al. (2022b). Machine learning for identifying benign and malignant of thyroid tumors: a retrospective study of 2,423 patients. *Front. Public Health* 10, 960740. doi:10.3389/fpubh.2022.960740

Gupta, P., Rustam, F., Kanwal, K., Aljedaani, W., Alfarhood, S., Safran, M., et al. (2024). Detecting thyroid disease using optimized machine learning model based on differential evolution. *Int. J. Comput. Intell. Syst.* 17 (1), 3. doi:10.1007/s44196-023-00388-2

Halder, A., Verma, A., Biswas, D., and Srivastava, S. (2021). Recent advances in mass-spectrometry based proteomics software, tools and databases. *Drug Discov. Today Technol.* 39, 69–79. doi:10.1016/j.ddtec.2021.06.007

Hu, A., Noble, W. S., and Wolf-Yadlin, A. (2016). Technical advances in proteomics: new developments in data-independent acquisition. *F1000Res* 5, 419. doi:10.12688/f1000research.7042.1

Huang, F., Tang, X., Ye, B., Wu, S., and Ding, K. (2022). PSL-LCCL: a resource for subcellular protein localization in liver cancer cell line SK_HEP1. *Database* 2022, baab087. doi:10.1093/database/baab087

Huang, F. Q., Li, J., Jiang, L., Wang, F. X., Alolga, R. N., Wang, M. J., et al. (2019a). Serum-plasma matched metabolomics for comprehensive characterization of benign thyroid nodule and papillary thyroid carcinoma. *Int. J. Cancer* 144 (4), 868–876. doi:10.1002/ijc.31925

Huang, L., Mao, X., Sun, C., Luo, Z., Song, X., Li, X., et al. (2019b). A graphical data processing pipeline for mass spectrometry imaging-based spatially resolved metabolomics on tumor heterogeneity. *Anal. Chim. Acta* 1077, 183–190. doi:10.1016/j.aca.2019.05.068

Huang, L., Song, M., Shen, H., Hong, H., Gong, P., Deng, H. W., et al. (2023). Deep learning methods for omics data imputation. *Biol. (Basel)* 12 (10), 1313. doi:10.3390/biology12101313

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., and Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics and Proteomics* 15 (1), 41–51. doi:10.21873/cgp.20063

Ivosev, G., Burton, L., and Bonner, R. (2008). Dimensionality reduction and visualization in principal component analysis. *Anal. Chem.* 80, 4933–4944. doi:10.1021/ac800110w

Jajin, M. G., Abooshahab, R., Hooshmand, K., Moradi, A., Siadat, S. D., Mirzazadeh, R., et al. (2022). Gas chromatography-mass spectrometry-based untargeted metabolomics reveals metabolic perturbations in medullary thyroid carcinoma. *Sci. Rep.* 12 (1), 8397. doi:10.1038/s41598-022-12590-x

Janitza, S., Tutz, G., and Boulesteix, A.-L. (2016). Random forest for ordinal responses: prediction and variable selection. *Comput. Statistics and Data Analysis* 96, 57–73. doi:10.1016/j.csda.2015.10.005

Jasem, N. M., Kadhim, S. H., and Sharba M, M. (2024). Assessing thyroid function: a review of biochemical markers and testing strategies. *Int. J. Med. Sci. Dent. Health* 10 (02), 130–150. doi:10.55640/ijmsdh-10-02-17

Jia, X., Zhai, T., and Zhang, J. A. (2021). Circulating exosome involves in the pathogenesis of autoimmune thyroid diseases through immunomodulatory proteins. *Front. Immunol.* 12, 730089. doi:10.3389/fimmu.2021.730089

Jimenez, C. R., and Verheul, H. M. W. (2014). Mass spectrometry-based proteomics:from cancer biology to protein biomarkers, drug targets, and clinical applications. *Am. Soc. Clin. Oncol. Educ. Book* 34, e504–e510. doi:10.14694/EdBook_AM.2014.34.e504

Johnson, K. B., Wei, W. Q., Weeraratne, D., Frisse, M. E., Misulis, K., Rhee, K., et al. (2021). Precision medicine, AI, and the future of personalized health care. *Clin. Transl. Sci.* 14 (1), 86–93. doi:10.1111/cts.12884

Joshi, N., Garapati, K., Ghose, V., Kandasamy, R. K., and Pandey, A. (2024). Recent progress in mass spectrometry-based urinary proteomics. *Clin. Proteomics* 21 (1), 14. doi:10.1186/s12014-024-09462-z

Jović, A., Brkić, K., and Bogunović, N. (2015). A review of feature selection methods with applications. *2015 38th Int. Convention Inf. Commun. Technol. Electron. Microelectron.*, 1200–1205. doi:10.1109/MIPRO.2015.7160458

Ross, Q. J. (1993). *C4.5: programs for machine learning*. San Francisco: Morgan Kaufmann Publishers.

Jiliang, T., Salem, A., and Huan, L (2014). Feature selection for classification:A review. *Data Classif. Algorithms Appl.* 37. doi:10.1201/b17320

Kang, M., Ko, E., and Mersha, T. B. (2022). A roadmap for multi-omics data integration using deep learning. *Brief. Bioinform* 23 (1), bbab454. doi:10.1093/bib/bbab454

Kappler, L., and Lehmann, R. (2019). Mass-spectrometric multi-omics linked to function – state-of-the-art investigations of mitochondria in systems medicine. *TrAC Trends Anal. Chem.* 119, 115635. doi:10.1016/j.trac.2019.115635

Karimpour-Fard, A., Epperson, L. E., and Hunter, L. E. (2015). A survey of computational tools for downstream analysis of proteomic and other omic datasets. *Hum. Genomics* 9 (1), 28. doi:10.1186/s40246-015-0050-2

Kavzoglu, T., and Teke, A. (2022). Predictive performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (XGBoost) and natural gradient boosting (NGBoost). *Arabian J. Sci. Eng.* 47 (6), 7367–7385. doi:10.1007/s13369-022-06560-8

Kawashima, Y., Watanabe, E., Umeyama, T., Nakajima, D., Hattori, M., Honda, K., et al. (2019). Optimization of data-independent acquisition mass spectrometry for deep and highly sensitive proteomic analysis. *Int. J. Mol. Sci.* 20 (23), 5932. doi:10.3390/ijms20235932

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., and King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 17 (1), 195. doi:10.1186/s12916-019-1426-2

Khalifa, M., and Albadawy, M. (2024). Artificial intelligence for diabetes: enhancing prevention, diagnosis, and effective management. *Comput. Methods Programs Biomed. Update* 5, 100141. doi:10.1016/j.cmpbup.2024.100141

Khalifa, M., Albadawy, M., and Iqbal, U. (2024). Advancing clinical decision support: the role of artificial intelligence across six domains. *Comput. Methods Programs Biomed. Update* 5, 100142. doi:10.1016/j.cmpbup.2024.100142

Khalilia, M., Chakraborty, S., and Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC Med. Inf. Decis. Mak.* 11 (51), 51. doi:10.1186/1472-6947-11-51

Kim, S., Kang, D., Huo, Z., Park, Y., Tseng, G. C., and Bar-Joseph, Z. (2018). Meta-analytic principal component analysis in integrative omics application. *Bioinformatics* 34 (8), 1321–1328. doi:10.1093/bioinformatics/btx765

Komuro, J., Kusumoto, D., Hashimoto, H., and Yuasa, S. (2023). Machine learning in cardiology: clinical application and basic research. *J. Cardiol.* 82 (2), 128–133. doi:10.1016/j.jjcc.2023.04.020

Kotsiantis, S. B. (2011). RETRACTED ARTICLE: feature selection for machine learning classification problems: a recent overview. *Artif. Intell. Rev.* 42 (1), 157. doi:10.1007/s10462-011-9230-1

Kowalczyk, T., Ciborowski, M., Kisluk, J., Kretowski, A., and Barbas, C. (2020). Mass spectrometry based proteomics and metabolomics in personalized oncology. *Biochim. Biophys. Acta Mol. Basis Dis.* 1866 (5), 165690. doi:10.1016/j.bbadis.2020.165690

Krassowski, M., Das, V., Sahu, S. K., and Misra, B. B. (2020). State of the field in multi-omics research: from computational needs to data mining and sharing. *Front. Genet.* 11, 610798. doi:10.3389/fgene.2020.610798

Krishnan, G., Singh, S., Pathania, M., Gosavi, S., Abhishek, S., Parchani, A., et al. (2023). Artificial intelligence in clinical medicine: catalyzing a sustainable global healthcare paradigm. *Front. Artif. Intell.* 6, 1227091. doi:10.3389/frai.2023.1227091

Kumari, P., Kaur, B., Rakhra, M., Deka, A., Byeon, H., Asenso, E., et al. (2024). Explainable artificial intelligence and machine learning algorithms for classification of thyroid disease. *Discov. Appl. Sci.* 6 (7), 360. doi:10.1007/s42452-024-06068-w

Kwon, M. R., Shin, J. H., Park, H., Cho, H., Hahn, S. Y., and Park, K. W. (2020). Radiomics study of thyroid ultrasound for predicting BRAF mutation in papillary thyroid carcinoma: preliminary results. *AJNR Am. J. Neuroradiol.* 41 (4), 700–705. doi:10.3174/ajnr.A6505

Leo, B., Jerome, F., and Olshe, R.A. (1984). *Classification and regression trees*. New York: Chapman and Hall, 582–588.

Leung Kwan, K. K., Wong, T. Y., Wu, Q. Y., Xia Dong, T. T., Lam, H., and Keung Tsim, K. W. (2021). Mass spectrometry-based multi-omics analysis reveals the thermogenetic regulation of herbal medicine in rat model of yeast-induced fever. *J. Ethnopharmacol.* 279, 114382. doi:10.1016/j.jep.2021.114382

Li, H., Ning, S., Ghandi, M., Kryukov, G. V., Gopal, S., Deik, A., et al. (2019). The landscape of cancer cell line metabolism. *Nat. Med.* 25 (5), 850–860. doi:10.1038/s41591-019-0404-8

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., et al. (2017a). Feature selection: a data perspective. *ACM Comput. Surv.* 50 (6), 1–45. doi:10.1145/3136625

Li, J., Mi, L., Ran, B., Sui, C., Zhou, L., Li, F., et al. (2023a). Identification of potential diagnostic and prognostic biomarkers for papillary thyroid microcarcinoma (PTMC) based on TMT-labeled LC-MS/MS and machine learning. *J. Endocrinol. Invest.* 46 (6), 1131–1143. doi:10.1007/s40618-022-01960-x

Li, S., Yi, H., Leng, Q., Wu, Y., and Mao, Y. (2023b). New perspectives on cancer clinical research in the era of big data and machine learning. *Surg. Oncol.* 52, 102009. doi:10.1016/j.suronc.2023.102009

Li, Y., Li, T., and Liu, H. (2017b). Recent advances in feature selection and its applications. *Knowl. Inf. Syst.* 53 (3), 551–577. doi:10.1007/s10115-017-1059-8

Li, Y., Wu, F., Ge, W., Zhang, Y., Hu, Y., Zhao, L., et al. (2024). Risk stratification of papillary thyroid cancers using multidimensional machine learning. *Int. J. Surg.* 110 (1), 372–384. doi:10.1097/js9.0000000000000814

Liang, L. W., Raita, Y., Hasegawa, K., Fifer, M. A., Maurer, M. S., Reilly, M. P., et al. (2022). Proteomics profiling reveals a distinct high-risk molecular subtype of hypertrophic cardiomyopathy. *Heart* 108 (22), 1807–1814. doi:10.1136/heartjnl-2021-320729

Liebal, U. W., Phan, A. N. T., Sudhakar, M., Raman, K., and Blank, L. M. (2020). Machine learning applications for mass spectrometry-based metabolomics. *Metabolites* 10 (6), 243. doi:10.3390/metabo10060243

Liu, J., Sun, W., Dong, W., Wang, Z., Qin, Y., Zhang, T., et al. (2017). HSP90 inhibitor NVP-AUY922 induces cell apoptosis by disruption of the survivin in papillary thyroid carcinoma cells. *Biochem. Biophys. Res. Commun.* 487 (2), 313–319. doi:10.1016/j.bbrc.2017.04.056

Lundberg, M. S., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Adv. neural Inf. Process. Syst.*, 4768–4777. doi:10.48550/arXiv.1705.07874

Luo, D., Zhan, S., Xia, W., Huang, L., Ge, W., and Wang, T. (2018). Proteomics study of serum exosomes from papillary thyroid cancer patients. *Endocr. Relat. Cancer* 25 (10), 879–891. doi:10.1530/ERC-17-0547

Ma, S., and Dai, Y. (2011). Principal component analysis based methods in bioinformatics studies. *Brief. Bioinform* 12 (6), 714–722. doi:10.1093/bib/bbq090

Manochkumar, J., Cherukuri, A. K., Kumar, R. S., Almansour, A. I., Ramamoorthy, S., and Efferth, T. (2023). A critical review of machine-learning for "multi-omics" marine metabolite datasets. *Comput. Biol. Med.* 165, 107425. doi:10.1016/j.compbiomed.2023.107425

Markowetz, F. (2001). *Support vector machines in bioinformatics* (Heidelberg: University of Heidelberg). *Master's thesis*.

Marsee, D. K., Venkateswaran, A., Tao, H., Vadysirisack, D., Zhang, Z., Vandre, D. D., et al. (2004). Inhibition of heat shock protein 90, a novel RET/PTC1-associated protein, increases radioiodide accumulation in thyroid cells. *J. Biol. Chem.* 279 (42), 43990–43997. doi:10.1074/jbc.M407503200

Martinez-Aguilar, J., Clifton-Bligh, R., and Molloy, M. P. (2015). A multiplexed, targeted mass spectrometry assay of the S100 protein family uncovers the isoform-specific expression in thyroid tumours. *BMC Cancer* 15, 199. doi:10.1186/s12885-015-1217-x

Martinez-Aguilar, J., Clifton-Bligh, R., and Molloy, M. P. (2016). Proteomics of thyroid tumours provides new insights into their molecular composition and changes associated with malignancy. *Sci. Rep.* 6, 23660. doi:10.1038/srep23660

Mavrogeorgis, E., He, T., Mischak, H., Latosinska, A., Vlahou, A., Schanstra, J. P., et al. (2023). Uniform manifold approximation and projection-based assessment of chronic kidney disease aetiologies based on urinary peptidomics. medRxiv. doi:10.1101/2023.05.19.23290228

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* 3 (29), 861. doi:10.21105/joss.00861

Mesarić, J., and Šebalj, D. (2016). Decision trees for predicting the academic success of students. *Croat. Operational Res. Rev.* 7 (2), 367–388. doi:10.17535/crorr.2016.0025

Arzyeh, G., Tristan, N., Peter, S., Andrew, L B., Irene, Y C., and Rajesh, R. (2020). A review of challenges and opportunities in machine learning for health. *AMIA Jt. Summits Transl. Sci.*, 191–200. doi:10.48550/arXiv.1806.00388

Migisha, N. P. V., Heo, C. E., Han, J. Y., Chae, S. Y., Kim, M., Vu, H. M., et al. (2020). Mass spectrometry-based proteomics of single cells and organoids: the new generation of cancer research. *TrAC Trends Anal. Chem.* 130, 116005. doi:10.1016/j.trac.2020.116005

Mirza, B., Wang, W., Wang, J., Choi, H., Chung, N. C., and Ping, P. (2019). Machine learning and integrative analysis of biomedical big data. *Genes. (Basel)* 10 (2), 87. doi:10.3390/genes10020087

Mischak, H., Allmaier, G., Apweiler, R., Attwood, T., Baumann, M., Benigni, A., et al. (2010). Recommendations for biomarker identification and qualification in clinical proteomics. *Sci. Transl. Med.* 2 (46), 46ps42. doi:10.1126/scitranslmed.3001249

Mohammadzadeh, B., Françoise, J., Gouiffès, M., and Caramiaux, B. (2024). "Studying collaborative interactive machine teaching in image classification," in Proceedings of the 29th International Conference on Intelligent User Interfaces, USA, January 13 - 16, 2004, 195–208. doi:10.1145/3640543.3645204

Mou, M., Pan, Z., Lu, M., Sun, H., Wang, Y., Luo, Y., et al. (2022). Application of machine learning in spatial proteomics. *J. Chem. Inf. Model.* 62 (23), 5875–5895. doi:10.1021/acs.jcim.2c01161

Mullur, R., Liu, Y. Y., and Brent, G. A. (2014). Thyroid hormone regulation of metabolism. *Physiol. Rev.* 94 (2), 355–382. doi:10.1152/physrev.00030.2013

Nalluri, M., Pentela, M., and Eluri, N. R. (2020). A scalable tree boosting system:XG boost. *Int. J. Res. Stud. Sci. Eng. Technol.* 7 (12), 36–51. doi:10.22259/2349-476X.0712005

Navada, A., Ansari, A. N., Patil, S., and Sonkamble, B. A. (2011). Overview of use of decision tree algorithms in machine learning. *IEEE Control Syst. Graduate Res. Colloquium*, 37–42. doi:10.1109/ICSGRC.2011.5991826

Ng, S., Masarone, S., Watson, D., and Barnes, M. R. (2023). The benefits and pitfalls of machine learning for biomarker discovery. *Cell. Tissue Res.* 394 (1), 17–31. doi:10.1007/s00441-023-03816-z

Ngan, H.-L., Lam, K.-Y., Li, Z., Zhang, J., and Cai, Z. (2023). Machine learning facilitates the application of mass spectrometry-based metabolomics to clinical analysis: a review of early diagnosis of high mortality rate cancers. *TrAC Trends Anal. Chem.* 168, 117333. doi:10.1016/j.trac.2023.117333

Nicholson, J. K., Lindon, J. C., and Holmes, E. (1999). 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* 29 (11), 1181–1189. doi:10.1080/004982599238047

Nipp, M., Elsner, M., Balluff, B., Meding, S., Sarioglu, H., Ueffing, M., et al. (2012). S100-A10, thioredoxin, and S100-A6 as biomarkers of papillary thyroid carcinoma with lymph node metastasis identified by MALDI imaging. *J. Mol. Med. Berl.* 90 (2), 163–174. doi:10.1007/s00109-011-0815-6

Omur, O., and Baran, Y. (2014). An update on molecular biology of thyroid cancers. *Crit. Rev. Oncology/Hematology* 90 (3), 233–252. doi:10.1016/j.critrevonc.2013.12.007

Pan, W., Shen, X., and Liu, B. (2013). Cluster analysis:unsupervised learning via supervised learning with a non-convex penalty. *J. Mach. Learn. Res.* 14 (7), 1865.

Paron, I., Scaloni, A., Pines, A., Bachi, A., Liu, F. T., Puppin, C., et al. (2003). Nuclear localization of Galectin-3 in transformed thyroid cells: a role in transcriptional regulation. *Biochem. Biophys. Res. Commun.* 302 (3), 545–553. doi:10.1016/s0006-291x(03)00151-7

Pearl, L. H., Prodromou, C., and Workman, P. (2008). The Hsp90 molecular chaperone: an open and shut case for treatment. *Biochem. J.* 410 (3), 439–453. doi:10.1042/BJ20071640

Perakakis, N., Yazdani, A., Karniadakis, G. E., and Mantzoros, C. (2018). Omics, big data and machine learning as tools to propel understanding of biological mechanisms and to discover novel diagnostics and therapeutics. *Metabolism* 87, A1-A9–A9. doi:10.1016/j.metabol.2018.08.002

Picard, D., Felsberg, J., Langini, M., Stachura, P., Qin, N., Macas, J., et al. (2023). Integrative multi-omics reveals two biologically distinct groups of pilocytic astrocytoma. *Acta Neuropathol.* 146 (4), 551–564. doi:10.1007/s00401-023-02626-5

Qian, L., Sun, R., Xue, Z., and Guo, T. (2023). Mass spectrometry-based proteomics of epithelial ovarian cancers: a clinical perspective. *Mol. Cell. Proteomics* 22 (7), 100578. doi:10.1016/j.mcpro.2023.100578

Qiu, S., Cai, Y., Yao, H., Lin, C., Xie, Y., Tang, S., et al. (2023). Small molecule metabolites: discovery of biomarkers and therapeutic targets. *Signal Transduct. Target Ther.* 8 (1), 132. doi:10.1038/s41392-023-01399-3

Quazi, S. (2022). Artificial intelligence and machine learning in precision and genomic medicine. *Med. Oncol.* 39 (8), 120. doi:10.1007/s12032-022-01711-1

Randall, L. V., Kim, D. H., Abdelrazig, S. M. A., Bollard, N. J., Hemingway-Arnold, H., Hyde, R. M., et al. (2023). Predicting lameness in dairy cattle using untargeted liquid chromatography-mass spectrometry-based metabolomics and machine learning. *J. Dairy Sci.* 106, 7033–7042. doi:10.3168/jds.2022-23118

Reel, P. S., Reel, S., Pearson, E., Trucco, E., and Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: a review. *Biotechnol. Adv.* 49, 107739. doi:10.1016/j.biotechadv.2021.107739

Remeseiro, B., and Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. *Comput. Biol. Med.* 112, 103375. doi:10.1016/j.compbiomed.2019.103375

Reska, D., Czajkowski, M., Jurczuk, K., Boldak, C., Kwedlo, W., Bauer, W., et al. (2021). Integration of solutions and services for multi-omics data analysis towards personalized medicine. *Biocybern. Biomed. Eng.* 41 (4), 1646–1663. doi:10.1016/j.bbe.2021.10.005

Ringner, M. (2008). What is principal component analysis? *Nat. Biotechnol.* 26 (3), 303–304. doi:10.1038/nbt0308-303

Nishimura, R., Yokose, T., and Mukai, K. (1997). S-100 protein is a differentiation marker in thyroid carcinoma of follicular cell origin:an immunohistochemical study. *Pathol. Int.* 47 (10), 673–679. doi:10.1111/j.1440-1827.1997.tb04440.x

Roca, C. P., Burton, O. T., Neumann, J., Tareen, S., Whyte, C. E., Gergelits, V., et al. (2023). A cross entropy test allows quantitative statistical comparison of t-SNE and UMAP representations. *Cell. Rep. Methods* 3 (1), 100390. doi:10.1016/j.crmeth.2022.100390

Rosenberger, F. A., Thielert, M., Strauss, M. T., Schweizer, L., Ammar, C., Madler, S. C., et al. (2023). Spatial single-cell mass spectrometry defines zonation of the hepatocyte proteome. *Nat. Methods* 20 (10), 1530–1536. doi:10.1038/s41592-023-02007-6

Sainburg, T., McInnes, L., and Gentner, T. Q. (2021). Parametric UMAP embeddings for representation and semisupervised learning. *Neural comput.* 33 (11), 2881–2907. doi:10.1162/neco_a_01434

Sarker, I. H. (2021). Machine learning: algorithms, real-world applications and research directions. *SN Comput. Sci.* 2 (3), 160. doi:10.1007/s42979-021-00592-x

Schneider, D. F., and Chen, H. (2013). New developments in the diagnosis and treatment of thyroid cancer. *CA Cancer J. Clin.* 63 (6), 374–394. doi:10.3322/caac.21195

Searle, B. C., Swearingen, K. E., Barnes, C. A., Schmidt, T., Gessulat, S., Kuster, B., et al. (2020). Generating high quality libraries for DIA MS with empirically corrected peptide predictions. *Nat. Commun.* 11 (1), 1548. doi:10.1038/s41467-020-15346-1

Singh, S., and Giri, M. (2014). Comparative study ID3, cart and C4. 5 decision tree algorithm:a survey. *Int. J. Adv. Inf. Sci. Technol. (IJAIST)* 27 (27), 97–103. doi:10.15693/ijaist/2014.v3i7.47-52

Smith, C. M., Weathers, A. L., and Lewis, S. L. (2023). An overview of clinical machine learning applications in neurology. *J. Neurological Sci.* 455, 122799. doi:10.1016/j.jns.2023.122799

Sofia, P., Sravani, T., and K Lakshmi, D. C. (2019). Anomalous development of thyroid gland; a cadaveric study in coastal population of Andhra Pradesh. *Indian J. Clin. Anat. Physiology* 6 (2), 220–223. doi:10.18231/j.ijcap.2019.049

Sofiadis, A., Dinets, A., Orre, L. M., Branca, R. M., Juhlin, C. C., Foukakis, T., et al. (2010). Proteomic study of thyroid tumors reveals frequent up-regulation of the Ca2+ -binding protein S100A6 in papillary thyroid carcinoma. *Thyroid* 20 (10), 1067–1076. doi:10.1089/thy.2009.0400

Speiser, J. L., Miller, M. E., Tooze, J., and Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Syst. Appl.* 134, 93–101. doi:10.1016/j.eswa.2019.05.028

Steven, J. (2017). Random forest. *J. Insur. Med.* 47 (1), 31–39. doi:10.17849/insm-47-01-31-39.1

Sugimoto, M., Kawakami, M., Robert, M., Soga, T., and Tomita, M. (2012). Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis. *Curr. Bioinforma.* 7 (1), 96–108. doi:10.2174/157489312799304431

Sun, Y., Salerno, S., He, X., Pan, Z., Yang, E., Sujimongkol, C., et al. (2023). Use of machine learning to assess the prognostic utility of radiomic features for in-hospital COVID-19 mortality. *Sci. Rep.* 13 (1), 7318. doi:10.1038/s41598-023-34559-0

Sun, Y., Selvarajan, S., Zang, Z., Liu, W., Zhu, Y., Zhang, H., et al. (2022). Artificial intelligence defines protein-based classification of thyroid nodules. *Cell. Discov.* 8 (1), 85. doi:10.1038/s41421-022-00442-x

Sun, Y., Wang, H., Li, L., Wang, J., Chen, W., Peng, L., et al. (2024). A diagnostic protein assay for differentiating follicular thyroid adenoma and carcinoma. medRxiv. doi:10.1101/2024.09.26.24314403

Surman, M., Wilczak, M., Jankowska, U., Skupien-Rabian, B., and Przybylo, M. (2024). Shotgun proteomics of thyroid carcinoma exosomes - insight into the role of exosomal proteins in carcinogenesis and thyroid homeostasis. *Biochim. Biophys. Acta Gen. Subj.* 1868 (9), 130672. doi:10.1016/j.bbagen.2024.130672

Tang, B., Pan, Z., Yin, K., and Khateeb, A. (2019). Recent advances of deep learning in bioinformatics and computational biology. *Front. Genet.* 10, 214. doi:10.3389/fgene.2019.00214

Torres-Cabala, C., Panizo-Santos, A., Krutzsch, H. C., Barazi, H., Namba, M., Sakaguchi, M., et al. (2004). Differential expression of S100C in thyroid lesions. *Int. J. Surg. Pathol.* 12 (2), 107–115. doi:10.1177/106689690401200203

Torun, F. M., Virreira Winter, S., Doll, S., Riese, F. M., Vorobyev, A., Mueller-Reif, J. B., et al. (2022). Transparent exploration of machine learning for biomarker discovery from proteomics and omics data. *J. Proteome Res.* 22 (2), 359–367. doi:10.1021/acs.jproteome.2c00473

Vanderpump, M. P. (2011). The epidemiology of thyroid disease. *Br. Med. Bull.* 99, 39–51. doi:10.1093/bmb/ldr030

Venkatesh, B., and Anuradha, J. (2019). A review of feature selection and its methods. *Cybern. Inf. Technol.* 19 (1), 3–26. doi:10.2478/cait-2019-0001

Wald, R., Khoshgoftaar, T. M., and Napolitano, A. (2013). "How the choice of wrapper learner and performance metric affects subset evaluation," in 2013 IEEE 25th International Conference on Tools with Artificial Intelligence, 4-6 Nov. 2013, 426–432. doi:10.1109/ictai.2013.70

Walsh, J. P. (2016). Managing thyroid disease in general practice. *Med. J. Aust.* 205 (4), 179–184. doi:10.5694/mja16.00545

Wang, F., Lin, Y., Xu, J., Wei, F., Huang, S., Wen, S., et al. (2024b). Risk of papillary thyroid carcinoma and nodular goiter associated with exposure to semi-volatile organic compounds: a multi-pollutant assessment based on machine learning algorithms. *Sci. Total Environ.* 915, 169962. doi:10.1016/j.scitotenv.2024.169962

Wang, G., Li, H. N., Cui, X. Q., Xu, T., Dong, M. L., Li, S. Y., et al. (2021). S100A1 is a potential biomarker for papillary thyroid carcinoma diagnosis and prognosis. *J. Cancer* 12 (19), 5760–5771. doi:10.7150/jca.51855

Wang, H., Zhang, C., Li, Q., Tian, T., Huang, R., Qiu, J., et al. (2024a). Development and validation of prediction models for papillary thyroid cancer structural recurrence using machine learning approaches. *BMC Cancer* 24 (1), 427. doi:10.1186/s12885-024-12146-4

Wang, X., Fan, D., Yang, Y., Gimple, R. C., and Zhou, S. (2023b). Integrative multi-omics approaches to explore immune cell functions: challenges and opportunities. *iScience* 26 (4), 106359. doi:10.1016/j.isci.2023.106359

Wang, X., Zhang, A., and Sun, H. (2013). Power of metabolomics in diagnosis and biomarker discovery of hepatocellular carcinoma. *Hepatology* 57 (5), 2072–2077. doi:10.1002/hep.26130

Wang, Y., Chen, Z., Shima, K., Zhong, D., Yang, L., Wang, Q., et al. (2022b). Rapid diagnosis of papillary thyroid carcinoma with machine learning and probe electrospray ionization mass spectrometry. *J. Mass Spectrom.* 57 (6), e4831. doi:10.1002/jms.4831

Wang, Y., Chen, Z., Zhang, L., Zhong, D., Di, J., Li, X., et al. (2022c). Fast classification of thyroid nodules with ultrasound guided-fine needle biopsy samples and machine learning. *Appl. Sci.* 12 (11), 5364. doi:10.3390/app12115364

Wang, Y., Lih, T. M., Chen, L., Xu, Y., Kuczler, M. D., Cao, L., et al. (2022a). Optimized data-independent acquisition approach for proteomic analysis at single-cell level. *Clin. Proteomics* 19 (1), 24. doi:10.1186/s12014-022-09359-9

Wang, Z., Wang, H., Zhou, Y., Li, L., Lyu, M., Wu, C., et al. (2024c). An individualized protein-based prognostic model to stratify pediatric patients with papillary thyroid carcinoma. *Nat. Commun.* 15 (1), 3560. doi:10.1038/s41467-024-47926-w

Wang, Z., Zhu, H., and Xiong, W. (2023a). Advances in mass spectrometry-based multi-scale metabolomic methodologies and their applications in biological and clinical investigations. *Sci. Bull. (Beijing)* 68, 2268–2284. doi:10.1016/j.scib.2023.08.047

Wekesa, J. S., and Kimwele, M. (2023). A review of multi-omics data integration through deep learning approaches for disease diagnosis, prognosis, and treatment. *Front. Genet.* 14, 1199087. doi:10.3389/fgene.2023.1199087

White, P. T., Subramanian, C., Zhu, Q., Zhang, H., Zhao, H., Gallagher, R., et al. (2016). Novel HSP90 inhibitors effectively target functions of thyroid cancer stem cell preventing migration and invasion. *Surgery* 159 (1), 142–151. doi:10.1016/j.surg.2015.07.050

Wickenberg, M., Mercier, R., Yap, M., Walker, J., Baker, K., and LaPointe, P. (2024). Hsp90 inhibition leads to an increase in surface expression of multiple immunological receptors in cancer cells. *Front. Mol. Biosci.* 11, 1334876. doi:10.3389/fmolb.2024.1334876

Wojakowska, A., Chekan, M., Marczak, L., Polanski, K., Lange, D., Pietrowska, M., et al. (2015). Detection of metabolites discriminating subtypes of thyroid cancer: molecular profiling of FFPE samples using the GC/MS approach. *Mol. Cell. Endocrinol.* 417, 149–157. doi:10.1016/j.mce.2015.09.021

Wojakowska, A., Cole, L. M., Chekan, M., Bednarczyk, K., Maksymiak, M., Oczko-Wojciechowska, M., et al. (2018). Discrimination of papillary thyroid cancer from non-cancerous thyroid tissue based on lipid profiling by mass spectrometry imaging. *Endokrynol. Pol.* 69 (1), 2–8. doi:10.5603/EP.a2018.0003

Ws, N. (2006). What is a support vector machine? *Nat. Biotechnol.* 24 (12), 1565–1557. doi:10.1038/nbt1206-1565

Xi, N. M., Wang, L., and Yang, C. (2022). Improving the diagnosis of thyroid cancer by machine learning and clinical data. *Sci. Rep.* 12 (1), 11143. doi:10.1038/s41598-022-15342-z

Xiao, J. F., Zhou, B., and Ressom, H. W. (2012). Metabolite identification and quantitation in LC-MS/MS-based metabolomics. *Trends Anal. Chem.* 32, 1–14. doi:10.1016/j.trac.2011.08.009

Xu, X., Fang, C., Wang, Y., Lu, F., and Liu, S. (2022). Integrating network pharmacology and metabolomics to elucidate the mechanism of action of Huang Qin decoction for treament of diabetic liver injury. *Front. Pharmacol.* 13, 899043. doi:10.3389/fphar.2022.899043

Yadav, S. (2024). Transformative frontiers: a comprehensive review of emerging technologies in modern healthcare. *Cureus* 16, e56538. doi:10.7759/cureus.56538

Yang, Y., Sun, H., Zhang, Y., Zhang, T., Gong, J., Wei, Y., et al. (2021). Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data. *Cell. Rep.* 36 (4), 109442. doi:10.1016/j.celrep.2021.109442

Yao, Z., Yin, P., Su, D., Peng, Z., Zhou, L., Ma, L., et al. (2011). Serum metabolic profiling and features of papillary thyroid carcinoma and nodular goiter. *Mol. Biosyst.* 7 (9), 2608–2614. doi:10.1039/c1mb05029j

Zhang, J., Feider, C. L., Nagi, C., Yu, W., Carter, S. A., Suliburk, J., et al. (2017a). Detection of metastatic breast and thyroid cancer in lymph nodes by desorption electrospray ionization mass spectrometry imaging. *J. Am. Soc. Mass Spectrom.* 28 (6), 1166–1174. doi:10.1007/s13361-016-1570-2

Zhang, S., Li, X., Zong, M., Zhu, X., and Cheng, D. (2017b). Learning k for kNN Classification. *ACM Trans. Intelligent Syst. Technol.* 8 (3), 1–19. doi:10.1145/2990508

Zhang, X., Lee, V. C., Rong, J., Lee, J. C., and Liu, F. (2022). Deep convolutional neural networks in thyroid disease detection: a multi-classification comparison by ultrasonography and computed tomography. *Comput. Methods Programs Biomed.* 220, 106823. doi:10.1016/j.cmpb.2022.106823

Zhang, Y., Liu, Y., Liu, H., and Tang, W. H. (2019). Exosomes: biogenesis, biologic function and clinical potential. *Cell. Biosci.* 9, 19. doi:10.1186/s13578-019-0282-2

Zhao, C., Dong, J., Deng, L., Tan, Y., Jiang, W., and Cai, Z. (2022a). Molecular network strategy in multi-omics and mass spectrometry imaging. *Curr. Opin. Chem. Biol.* 70, 102199. doi:10.1016/j.cbpa.2022.102199

Zhao, F., Zhang, H., Cheng, D., Wang, W., Li, Y., Wang, Y., et al. (2022b). Predicting the risk of nodular thyroid disease in coal miners based on different machine learning models. *Front. Med. (Lausanne)* 9, 1037944. doi:10.3389/fmed.2022.1037944

Zhou, J., Li, Y., Chen, X., Zhong, L., and Yin, Y. (2017). Development of data-independent acquisition workflows for metabolomic analysis on a quadrupole-orbitrap platform. *Talanta* 164, 128–136. doi:10.1016/j.talanta.2016.11.048

Zhu, X., Liu, Y., Tang, X., Sun, Y., Yi, H., Wang, J., et al. (2024). Feature screening of radioactive iodine-refractory thyroid carcinoma based on proteomics analysis and artificial intelligence. *SSRN*. doi:10.2139/ssrn.4865048