Metabolic profiling on 2D NMR TOCSY Spectra Using Machine Learning

Von der Fakultät
Elektrotechnik und Informationstechnik
der Technischen Universität Dortmund
genehmigte

Dissertation

zur Erlangung des akademischen Grades
Doktor der Naturwissenschaft (Dr. rer. nat.)

eingereicht von

Lubaba Yousef Hazza Migdadi, M.Sc.

Dortmund, 2023

Tag der mündlichen Prüfung: 02.06.2023

Hauptreferent: Prof. Dr. rer. nat. habil. Christian Wöhler, TU Dortmund

Korreferent Prof. Dr.-Ing. habil. Franz Kummert, Bielefeld University

---

Arbeitsgebiet Bildsignalverarbeitung Technische Universität Dortmund

Abstract

Metabolomics is an expanding field of medical diagnostics due to metabolic reprogramming alteration caused through diseases. Additionally, studying metabolomics offers an insight into the molecular mechanisms of diseases. The dynamicity of biological cells causes alteration in the chemical and biochemical characteristics of structural profiles of biological fluids and tissues. Therefore, the role of metabolic profiling in discovering biological fingerprints of diseases, and their evolution, as well as the cellular pathway of different biological or chemical stimuli is most significant.

Two-dimensional nuclear magnetic resonance (2D NMR) is one of the fundamental and strong analytical instruments for metabolic profiling. Though, total correlation spectroscopy (2D NMR ^{1}H -^{1}H TOCSY) can be used to improve spectral overlap of 1D NMR, strong peak shift, signal overlap, spectral crowding and matrix effects in complex biological mixtures are extremely challenging in 2D NMR analysis. Thus, in this work, we introduce an automated metabolic deconvolution and assignment based on TOCSY of real breast cancer tissue and of adipose tissue-derived human Mesenchymal Stem cells. A major alternative to the common approaches in NMR based machine learning where images of the spectra are used as an input. In the new suggested approach, metabolic assignment is based only on the vertical and horizontal frequencies of the metabolites in the ^{1}H-^{1}H TOCSY.

A set of 27 metabolites were deduced from the TOCSY of a breast cancer sample and the classifiers: Kernel Null Foley–Sammon Transform, support vector machines, and third- and fourth-degree polynomial classifiers have been customized and extended under the semi-supervised learning scheme. The classifiers' performance was evaluated by comparing the conventional human-based methodology and automatic assignments under different initial training sizes settings.

Most metabolic profiling approaches focus only on identifying pre-known metabolites on ^{1}H-^{1}H TOCSY spectrum using configured parameters. However, there is a lack of research dealing with automating the detection of new metabolites that might appear during the dynamic evolution of biological cells. Novelty detection is a category of machine learning that is used to identify data that emerge during the test phase and were not considered during the training phase. We propose a novelty detection system for detecting novel metabolites in the 2D NMR ^{1}H-^{1}H TOCSY spectrum of a breast cancer-tissue sample. We build one- and multi-class recognition systems using different classifiers such as Kernel Null Foley-Sammon Transform, Kernel Density Estimation, and Support Vector Data Description. The training models were constructed based on different sizes of training data and are used in the novelty detection procedure. Multiple

evaluation measures were applied to test the performance of the novelty detection methods. The results of our novel metabolic profiling method demonstrate its suitability, robustness, and speed in automated metabolic research.
Furthermore, machine learning is applied on real-time 2D ^{1}H-^{1}H TOCSY to monitor the dynamic evolution of adipose tissue-derived human Mesenchymal Stem cells (AT-derived hMSCs) cultivated in basal culture media or in the presence of adipogenic or osteogenic differentiation media for a duration of fourteen days. Multi-class classification in addition to novelty detection of metabolites were established based on the profile of control hMSCs sample at four days cultivation and successively detect the absence and the abundance of metabolites after fourteen days of cultivation, adiobocytes and osteocytes differentiation. Kernel Null Foley-Sammon Transform und Kernel Density Estimation were successfully able to reveal metabolic changes that accompany MSCs cellular evolution starting from the undifferentiated status to their prolonged cultivation and differentiation into adipocytes and osteocytes. The results show high performances of the proposed algorithms and are compatible with the proved scientific analysis in stem cells differentiation studies.

Table of Contents

Abstract iii
Table of Contents v
Publications and Datasets vii
Dedication viii
Acknowledgment ix
1. Introduction 1
1.1. Motivation 1
1.2. Contributions 3
1.3. Thesis Outline 4
2. Machine Learning 7
2.1. Introduction 7
2.2. Supervised Learning 8
2.3. Unsupervised Learning 9
2.4. Semi-supervised Learning 9
2.5. Confidence bands 13
2.6. Novelty Detection 13
2.7. Threshold Computation 16
3. Nuclear Magnetic Resonance 18
3.1. Nuclear Magnetic Resonance (NMR) 18
3.2. One-Dimensional Nuclear Magnetic Resonance 18
3.3. Two-Dimensional NMR spectroscopy (2D NMR) 20
4. NMR-Based Metabolomics 25
4.1. 1D NMR Metabolite Spectra 25
4.2. 2D NMR Metabolite Spectra and Metabolite Assignment 26
4.3. Automated Metabolite Identification 27
4.4. Related Work 28
5. Datasets 32
5.1. NMR data acquisition and processing 32
5.2. TOCSY crosspeak picking and de-noising 32
5.3. Breast cancer tissue cells 33
5.4. Adipose tissue-derived human Mesenchymal Stem cells 36
5.5. Data Representation 37

6. Contribution: Semi-Supervised Learning in Metabolomics employing 2D TOCSY Spectrum 44
6.1. Semi-supervised Polynomial Classifier 45
6.2. Semi-supervised Support Vector Machines 45
6.3. Semi-supervised Kernel Null Foley–Sammon Transform 47
6.4. Experiments 51
6.5. Results and Discussion 54
6.6. Validation 56
6.7. Conclusion 60
7. Contribution: Novelty Detection in Metabolomics Employing 2D TOCSY Spectrum 62
7.1. Kernel Null Foley-Sammon transform 63
7.2. Support Vector Data Description 63
7.3. Kernel Density Estimation 65
7.4. Threshold setting and novelty detection 65
7.5. Novelty detection of metabolites using breast cancer tissue 66
7.6. One-class novelty detection 67
7.7. Multi-class novelty detection 71
7.8. Conclusions 74
8. Contribution: Automated Monitoring of Metabolic Changes Accompanying the Differentiation of Adipose Tissue-Derived Human Mesenchymal Stem Employing ^{1}H-^{1}H TOCSY NMR 75
8.1. Machine learning 76
8.2. Metabolic evolution of AT-derived hMSCs 76
8.3. Conclusion 82
9. Summary and Conclusions 83
10. Appendix 86
A. Novelty detection related results 86
B. AT-derived hMSCs Sample preparation 93
List of Figures 96
List of Tables 98
List of Acronyms 99
Bibliography 101

Publications and Datasets

Peer-reviewed journal articles

This thesis is based on the following publications of the author. The publications are listed in ascending chronological order.

1. Migdadi, L.; Lambert, J.; Telfah, A.; Hergenröder, R.; Wöhler, C., Automated metabolic assignment: Semi-supervised learning in metabolic analysis employing two dimensional Nuclear Magnetic Resonance (NMR). Computational and Structural Biotechnology Journal 2021, 19, 5047-5058. https://doi.org/10.1016/j.csbj.2021.08.048

2. Migdadi, L.; Telfah, A.; Hergenröder, R.; Wöhler, C., Novelty detection for metabolic dynamics established on breast cancer tissue using 2D NMR TOCSY spectra. Computational and Structural Biotechnology Journal 2022, 20, 2965-2977. https://doi.org/10.1016/j.csbj.2022.05.050

3. Migdadi, L.; Sharar, N.; Jafar, H.; Telfah, A.; Hergenröder, R.; Wöhler, C., Machine Learning in Automated Monitoring of Metabolic Changes Accompanying the Differentiation of Adipose-Tissue-Derived Human Mesenchymal Stem Cells Employing ^{1}H-^{1}H TOCSY NMR. Metabolites 2023, 13 (3), 352. https://doi.org/10.3390/metabo13030352

Datasets

1. Migdadi, L.; Lambert, J.; Telfah, A.; Hergenröder, R.; Wöhler, C. (2021). Automated metabolic assignment: Semi-supervised learning in metabolic analysis employing two dimensional Nuclear Magnetic Resonance (NMR) [Data set]. https://doi.org/10.5281/zenodo.5724057

2. Migdadi, L.; Sharar, N.; Jafar, H.; Telfah, A.; Hergenröder, R.; Wöhler, C. (2022). Dataset for publication: Machine Learning in Automated Monitoring of Metabolic Changes Accompanying the Differentiation of Adipose Tissue-Derived Human Mesenchymal Stem cells employing ^{1}H-^{1}H TOCSY NMR [Data set]. https://doi.org/10.5281/zenodo.7276518

Dedication

To the Loving Memory of My Father, Dr. Yousef Migdadi.
To My Marvelous Mother, Ms. Rowaida Alomari.

Acknowledgment

First and foremost, I would like to express my sincere thanks to my supervisor Prof. Dr. Christian Wöhler for giving me the opportunity to pursue a PhD in his department. I am extremely fortunate to have the ability to work closely with Prof. Wöhler and really appreciate his unlimited support and thoughtfulness during the past years. Thank you for your trust and confidence in my capability to advance my PhD and in my ability to succeed in my work. Without his treasured weekly meetings, suggestions, guidance, and supervision, I would not have been able to accomplish my tasks. I would also like to thank my colleagues at the image analysis group at TU-Dortmund for their valuable discussions during the weekly meetings.

I am in-depth grateful to Dr. Roland Hergenröder from ISAS for agreeing to start this project at ISAS. It was indeed a great opportunity and pleasure working in the Bioresponsive Materials group under his leadership. His unconstrained managing style and enthusiasm for the topics I worked on encouraged me to accomplish my work. In addition, I really appreciate providing a perfect working environment on the two DAAD-projects I have been coordinating. I am also appreciative to my colleagues in ISAS, Ahmad Bahti, Mais Ahmad, and Qais Al Bataineh for the relaxing hours we spent together and the scientific discussions we had.

I am deeply grateful for my extraordinary mother, sisters, and brothers for their unconditional support and tolerance during challenging times. My late father would have been immensely proud and elated at my achievements. I hold in high regard the values and principles my parents instilled in us. I dedicate this work to their memory.

Ahmad's role in both my scientific and personal life cannot be adequately expressed in words. He is not only my husband but also a colleague, a partner, a supporter, and a friend who offers unconditional love, trust, faith, and care. Meeting Ahmad was the most precious triumph of my life.

I cannot begin to express my gratitude and appreciation to my precious daughters for their patience, endurance, and their lovely, delightful comments. I am forever indebted to them.

This dissertation bears as evidence of the assistance, support, encouragement, and absolute belief of my family in me.

Thank you!

Lubaba

No amount of observations of white swans can allow the inference that all swans are white, but the observation of a single black swan is sufficient to refute that conclusion.

David Hume

1. Introduction

1.1. Motivation.....1
1.2. Contributions.....3
1.3. Thesis Outline.....4

1.1. MOTIVATION

Metabolomics is defined as the "the quantitative measurement of the multi-parametric metabolic response of living systems to pathophysiological stimuli or genetic modification" [1]. In organisms, metabolites are in dynamic interaction within body cells, tissues, and environment. As a result, any biological alteration in the regular cellular process in the body will be revealed in an alteration of biofluid composition. These alterations are considered as biomarkers or biological signature that could expose the characteristics of the biochemical status [1, 2]. Altered metabolism, sometimes called 'metabolic reprogramming,' caused by diseases offers an insight into the molecular mechanisms of diseases. This provides a sound basis for the identification of diagnostic and prognostic biomarkers, tracking diseases development and treatment outcomes as well as for rational drug design [3]. Even at initial stages, tumors have been found to modify the metabolic profiles of biofluids like e.g., blood and urine, as well as of tissues, resulting in fluctuations of the concentrations of already existing markers or in the generation of new ones. Consequently, metabolomics and metabolic profiling are considered a promising area that involve the detection and the identification of the biomarkers related to prognosis and diagnosis of biological abnormalities [4].

There is a demanding necessity of developing distinctive bioinformatics methods for metabolic identification due to the following challenges in metabolomics. First, the diversity, dynamicity and the complexity of the metabolites that can be found in a living system introduces an extra complication in metabolic analysis. In addition, absorption, synthesis, degradation and interaction with the environment are continuous processes that cause instant changes of the metabolism [5, 6]. Consequently, a distinct metabolomics profile that reveals the state of disease and the essential organism characteristics can be recognized, enabling further improvements in the diagnostic and prognostic methods ,and the detections of abnormal metabolic connection [7]. Moreover, metabolomics studies different types of chemical pathways, such as acids or lipids which further complicates the analysis process [6, 8]. Furthermore, metabolites have strong correlations between variables and in NMR one metabolite can contribute to multiple signals and different metabolites are connected through physiological pathways which

adds to the complexity of metabolites identification [9]. Additionally, metabolomics concentrate on downstream outcomes of organisms [6]; thus, the metabolome reflects the true dynamic functional state of cells and acts as explicit signatures of biochemical interactions and responses to genetic or environmental changes [7, 10]. Therefore, it is vital to choose analytical methods for the purpose of identification of diagnostic biomarkers allowing for further processing and analysis of the biological samples.

Nuclear magnetic resonance (NMR) spectroscopy is a powerful technique for the identification of the components of complex mixtures of small molecules, e.g., metabolites. NMR has proven its vital and powerful role as an analytical technique in metabolomics. The non-destructiveness and the reproducibility of NMR results lead to enabling high-throughput identification and quantitative accuracy of the metabolic concentration in biological mixtures [3, 11]. However, due to the low sensitivity and resolution in NMR, obtaining metabolic profiling data from NMR spectra is one of the main challenges in analyzing complex biological mixtures. Low sensitivity and resolution in NMR lead to signal overlapping in a ^{1}H NMR spectrum and metabolites are effected by peak shift due to pH and ionic strength variations of the biological matrix of the measured samples [3, 12, 13]. Therefore, consistent metabolic identification in biological fluids such as blood and urine or tissue [11] from the 1D NMR spectra is one of the significant challenges since it requires deconvolution of the NMR spectrum to overcome the spectral superposition of several metabolites [13, 14]. In principle, metabolic identification might be achieved by separating the mixture components by physical means, followed by NMR measurements of each component. In this approach, the overall NMR spectrum is assumed to correspond to a weighted sum of individual metabolite spectra measured individually or taken from an available reference dataset. Accordingly, concurrent metabolic identification by accurately matching the measured metabolites in the sample with the peak positions of the reference spectra can be achieved [14]. This approach is performed manually and involves considerable experience in NMR spectroscopy, metabolic assignment, sample type and chemical structure and is prone to operator bias [13, 14]. Moreover, this procedure is not only time-consuming, labor-intensive, and impractical but might also be invasive since some metabolites may lose their activity during separation [6]. Therefore, samples are measured without chemical separation into individual metabolites, and afterward, the deconvolution of the resulting NMR spectrum is performed based on specific approaches such as "targeted metabolite fitting" [14-16]. Fortunately, in many cases, peaks that overlap in 1D NMR spectra can be resolved in 2D NMR spectra due to their higher spectral dispersion [11, 17]. Therefore, ^{1}H-^{1}H TOCSY (total correlation spectroscopy) is well suited for spectral dispersion. Consequently, metabolomics assignments can be achieved as the signals of each metabolite occur on a single line (1D cross-sections (row) in the TOCSY spectrum). This approach eases the task of assignment as well as computational analysis. Nevertheless, automatically analyzing metabolites contained in biological mixtures using TOCSY spectra is currently limited [11]. Although many existing methods can decompose the mixed-signal spectrum into the individual spectra of the constituent metabolites, they cannot cope with the presence of spectral components induced by chemical shifts and

overlapping of metabolites. Therefore, the above-mentioned issues faced during analyzing 1D NMR are valid for 2D NMR. Concisely, the manual analysis of biological applications is considered a major challenge for high-throughput experiments, due to complexity of the experimental results [18] and the shortage of experts [19]. NMR chemical shift automatic assignment boosted by the ability of detecting new unexpected metabolites will offer a comprehensive characterization of the dynamic changes of metabolites, and the functional relationship in the metabolic pathways [20]. Machine learning and pattern recognition have been recognized as an important method for automation the drug discovery [21], analysis of bio systems such as enzymes, pathways, and cells biology [22, 23], in addition to structural and system biology [24].

1.2. CONTRIBUTIONS

Machine learning appears as a compelling development in NMR spectroscopic metabolic profiling. We establish automated metabolic assignment systems based on the spectral deconvolution of 2D TOCSY NMR by employing machine learning models. Multiple classifiers are built and optimized for automatic metabolite assignment of different biological samples under different training dataset sizes. Moreover, a database of metabolites was constructed through utilizing the horizontal and vertical frequencies of the TOCSY spectra. This metabolic database has been used in our system and can be further employed and updated for future metabolic assignment tasks. The results of the automated procedures are compared to manual analysis by experts. The contributions of this work are:

1. Semi-Supervised Learning (SSL) in metabolomics employing 2D TOCSY Spectra: SSL is implemented to assign labels to the different peaks in the TOCSY spectrum. SSL is helpful in cases where shortage of already existing training labeled data is encountered. SSL uses a combination of the already labeled data and the unlabeled data to assign the peaks to specific metabolites. The quality of the automated labelling is tested using an independent data set.
2. Novelty Detection (ND) in metabolomics employing 2D TOCSY Spectra: Due to the dynamic nature of biological cells and the variability and multifaceted corresponding biochemical responses, discovery of unexpected novel biomarker which may emerge due to an internal or external stimuli is substantial. Distinguishing these biomarkers is essential in drug design, personalized therapy and understanding the biological pathway and the biochemical mechanisms of recovery and degeneration.
3. Automated monitoring of metabolic changes accompanying the differentiation of Adipose tissue-derived human mesenchymal stem cells (AT-derived hMSCs) employing ^{1}H-^{1}H TOCSY NMR: a real monitoring of the differentiation of AT-derived hMSCs to identify the metabolic pathways through different types of differentiations and long cultivation is studied and compared to established studies related to stem cells differentiation.

Most of the modern automated tools that are employed to analyze TOCSY spectra use images of the spectrum as an input to neural networks or use multivariate statistical

analysis, such as Principal Component Analysis (PCA) and least squares method for the purpose of classification [25]. On the other hand, a significant emphasize in the methods described in this work is incorporating the frequencies of the TOCSY spectra in the assignment process. The usage of frequencies instead of images has the following advantages. Frequencies are directly related to the chemical shift values (ppm) values. PPM is a representation of characteristic frequency of the NMR device with respect to standard reference point and is independent from the spectrometer frequency, therefore, they can be adapted according to the frequency of the NMR spectrometer. These values acts like a fingerprint of a nuclei in biological components [26]. Moreover, ppm values are easily accessible, are standardized in unified databases, and are consistent and reproducible under predefined protocols [11, 27]. On the other hand, images of TOCSY spectra are inherently noisy as can be seen in Figure 1.1 [28] and are dependent on the measurement resolution and sensitivity. Our Noise Suppression procedure is discussed in Chapter 5.

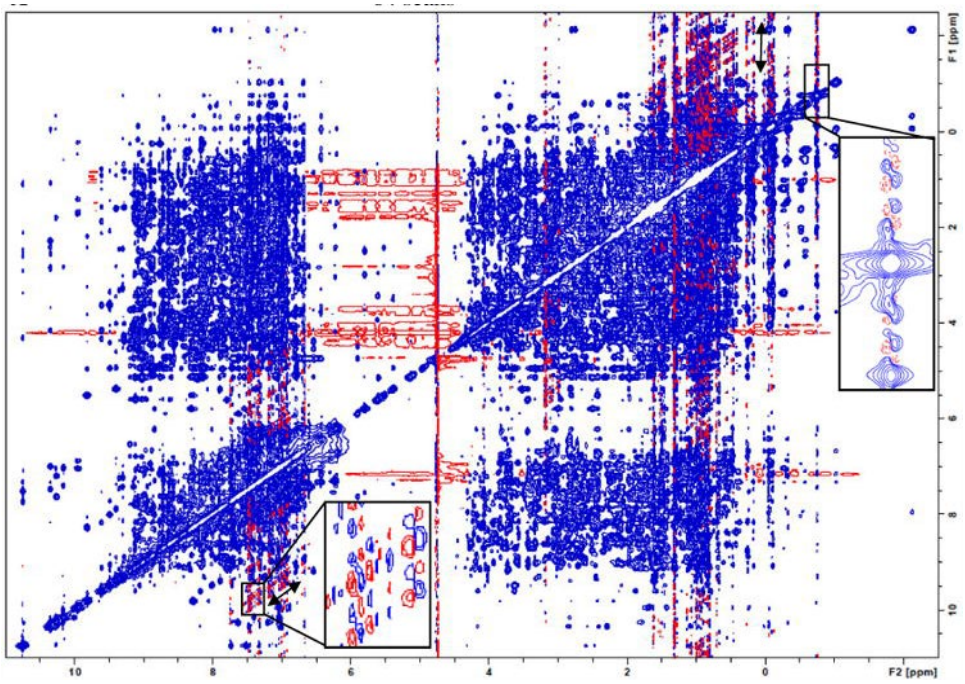


Figure 1.1: A noisy 2D NMR spectrum. Especially for samples with low intensity, NMR signal is contaminated by noise which appears as random fluctuating streaks in 2D NM resulting in reduced spectrum quality [28].

1.3. THESIS OUTLINE

The thesis is structured as follows. Consequent to the introduction, Chapter 2 introduces basic machine learning principles, important terms and concepts of semi-supervised learning and novelty detection are introduced, and a summary of different methods are given. In Chapter 3, the general foundation of nuclear magnetic resonance (NMR) spectroscopy and the principles of 1D and 2D NMR are presented. In Chapter 4 the significant role of NMR in metabolic studies in addition to the importance of automating the metabolic assignment in NMR analysis are discussed. Additionally, relevant

contributions and related work that employ machine learning and NMR are covered. In Chapter 5, NMR data acquisition, dataset construction and experimental setup are discussed. Semi-supervised learning techniques are implemented using 2D NMR TOCSY spectra of breast-cancer tissue samples in Chapter 6. While the assignment of known metabolites is the topic of Chapter 6, the detection of novel metabolites in 2D NMR TOCSY spectra is conducted in Chapter 7. An experiment that simulates the metabolic changes in metabolism of breast-cancer tissue sample was designed to test the performance of the classifiers. In Chapter 8, monitoring of metabolic pathways of Adipose tissue-derived human MSCs (AT-derived hMSCs) cultivated in basal culture media or in the presence of adipogenic or osteogenic differentiation media for a duration of fourteen days was conducted. Chapter 9 concludes the thesis and offers potential future research extension. Related data and results are presented in the Appendix.

2. Machine Learning

2.1. Introduction....7
2.2. Supervised Learning....8
2.3. Unsupervised Learning....9
2.4. Semi-supervised Learning....9
2.5. Confidence bands....13
2.6. Novelty Detection....13
2.7. Threshold Computation....16

Metabolic profiling of NMR spectra of biofluid samples and tissues are affected by extreme peaks shift and peaks overlap which lead to spectral crowding [13, 14]. Spectral crowding hardens the process of peak identification, multiplicity and J-couplings determination in addition to structural investigation [5]. Consequently, manual assignment of NMR spectra of complex mixtures is a tedious, time and labor-intensive task and depends extremely on expert knowledge [1, 29]. Developing an automatic system for peaks assignment of NMR spectra is of significant importance [11]. In this chapter, an overview of the machine learning methodologies used in this work is given.

2.1. INTRODUCTION

Machine learning has been defined by Tom Mitchell [30] as "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks T, as measured by P, improves with experience E".
The vital principle of machine learning is that it concentrates on the utilizing of procedures that incorporates information related to training data (experience E) to automatically estimate the model parameters, generalize to new data and make predictions (tasks T) to increase accuracy of classification (performance measure P) [30].
Machine learning is offering innovative insights into our lives. Every day, individuals interact with machine learning centered systems. For instance, voice recognition systems in intelligent personal assistant like Alexa and Google Assistant; e-mail spam filtering; image and face recognition in smart phones, security applications and social networks; weather prediction; traffic and map applications and customer service in retail applications. Machine learning is an emerging technology in many fields due to its effectiveness and scalability across a suite of applications. Machine learning offers a competitive advantage due to the high computational power of modern technology which enables the use of high computational resources and the integrating, collecting and organizing of large amount of data [31, 32].
In biology and medicine, machine learning is supporting scientists in prediction evaluation, uncertainty estimation and model interpretation methods of medical images,

including x-ray, MRI, and mammography images. Further applications in life science include disease and patients categorization using molecular biomarkers; enabling the utilizing of data from high-throughput 'omics' including genomics, proteomics, or metabolomics; recommendation of treatment, predicting treatment results and the development of new medications [33].

In general, a machine learning system uses three types of datasets: First dataset is the training dataset which is the labeled training data used to build a generalization model. The second dataset is the learning dataset which is the unlabeled data that is to be learned. A third dataset, the validation dataset, is used to tune the parameters of the classifiers. Importantly, all datasets must belong to the same distribution, but, the learning dataset is still unknown to the classifier during the training phase [34, 35].

Murphy [35] has divided machine learning into two main categories, supervised and unsupervised learning. These categories imply different amounts of supervision which is reflected in how much information is shared from the human expert side. Equally, the amount of information shared from the human expert side binds our choice of the chosen machine-learning category. Supervised learning corresponds to finding a mapping using the labeled training data for the purpose of assigning labels to unlabeled data. Usually, labeled training data are labeled by humans and is available before starting the learning procedure [35]. On the contrary, in unsupervised learning, classifiers receive a completely unlabeled dataset. The machine learning system is supposed to create clusters or groups based on similarities or a hidden structure in the training data [34, 36].

Usually, experts assign labels manually; nevertheless, having a complete set of manually labeled training data is a challenging task. The manual labelling process is time-consuming, depends extremely on expert's knowledge and is inflexible in cases of many unlabeled data from high-throughput applications. In these situations, a third category of machine learning that combines the previously mentioned categories is incorporated. Semi-supervised learning uses few labeled and many unlabeled data to infer the learning behavior and increase the classification performance. The labeled training data acts as seeds to create an initial training model; then the training model and the unlabeled data are used to update the initial training model [37].

Another emerging category is novelty detection. Novelty detection is defined as distinguishing test samples that differ from the training data [38]. Novelty detection is used when training data is incomplete or a particular class happens very rarely or in abnormal situations [39]. A brief introduction of these categories is introduced in the following sections.

2.2. SUPERVISED LEARNING

The goal of supervised learning is to learn the function that can predict the label of the unlabeled unseen instance. Formally, we have a training dataset $T = \{(x_i, y_i)\}_{i=1}^{N}$ where N is the number of training samples. Every input-output pair comprises a training input, $x_i \in X$, where X is a D-dimensional vector that represents the features or the attributes of the N training instance, together with its output label y_i. In cases of continuous or real-scaled label values, the prediction task is known as regression. On the other hand, in cases

when the prediction output $y \in \{1, \dots, C\}$ is assumed to be categorical or belong to discrete classes C, the prediction task is termed as classification. Classification can be regarded as a binary classification problem if $C = 2$ or as multi-class classification [35] if $C > 2$.
Supervised learning can be formulated as a generative or discriminative model. The generative algorithms learn the joint probability distribution $p(x, y)$ to model the class distributions. On the other hand, discriminative models learn the boundary between classes directly using the training data by estimating the posterior class probabilities $p(y|x)$ without modelling the underlying class distribution [35, 40] .
For metabolic assignment using supervised learning, a classifier must be trained over an interval of possible shifted frequencies and over intervals of possible pH values, concentrations, temperature, and any other effect that may affect the chemical shift. Unfortunately, supervised learning cannot be used efficiently in NMR experiments due to the inapplicability to capture all settings in the dynamic environment of metabolites.

2.3. UNSUPERVISED LEARNING

In unsupervised learning, no labelling information is available for the classifier and only the unlabeled training data X is used during the training process. Unsupervised learning is used to create clusters or groups of structures based on the similarities between data. In general, unsupervised learning is used to estimate the probability $p(x_i)$ under the assumption that $x_i \in X$ is independent and identically distributed [34].
Unsupervised learning will not be studied in this thesis.

2.4. SEMI-SUPERVISED LEARNING

One of the main problems in supervised learning is the unavailability of labeled training data. The number of training data samples, which are sufficient to produce an acceptable classification results, is directly related to the complexity of the classification problem [41]. Conversely, manual labeling is an expensive and time-consuming process, which needs a considerable amount of human supervision. In addition, in some context; the output of labeling process varies depending on the experience of the expert and therefore prone to error [37].
Semi-Supervised Learning (SSL) is a category that lies between supervised and unsupervised learning. Some paradigms view SSL as an extension of supervised learning with some extra information. This view is acceptable when the goal of the classification process is assigning labels. SSL can also be viewed as extended unsupervised learning with some constrains on the construction of clusters [41].
In SSL, a system is provided with a limited amount of labeled training data $\mathrm{X_{labeled}} = \{x_1, \dots, x_L\}$ and the associated labels $\mathrm{Y_{labeled}} = \{y_1, \dots . y_L\}$ drawn from $p(x, y)$ and unlabeled training data $\mathrm{X_{unlabeled}} = \{x_{L+1}, \dots, x_{L+u}\}$ from the marginal distribution $p(x)$, where we have L labeled data and u unlabeled data and $u \gg L$. SSL aims to use both $\mathrm{X_{unlabeled}}$ and $\mathrm{X_{labeled}}$ to boost the performance of the supervised learning based only on $\mathrm{X_{labeled}}$ [37]. SSL is based on at least one of these assumptions [42, 43]:

- Smoothness Assumption: in SSL, smoothness is related to density on the decision boundaries. Close instances in the input space are likely to belong to the same class,

constructing a high-density area. Accordingly, two instances of different classes are separated by low-density regions.

- Cluster Assumption or low-density assumption: decision boundary lies in low density regions and encloses high-density areas.
- Manifold Assumption: high dimensional data can be embedded in low dimensional area and can be handled in this low dimensional manifold. This assumption implies that the high dimensional input space can be represented through collection of low-dimensional manifolds that contains data instances and every manifold represents similar classes [42, 43].

The smoothness and the cluster assumptions are closely related. Clusters are formed using similar data, these data are grouped together to create a high-density area. The boundaries between these clusters form low density regions, which distinguish or separate these clusters [42]. The manifold assumption is a vital assumption in the dimensionality reduction of high dimensional features. In real applications, finding a low dimensional representation, which preserve the non-linear high dimensional information of the data, is of interest [37, 44].
Most SSL algorithms are derived from one of the above-mentioned assumptions; in correspondence to the underlying assumption, SSL is organized as follows [37, 41, 44, 45]:

2.4.1. Graph-based methods

In Graph-based SSL, labeled and unlabeled data form vertices, which are connected through weighted edges. The edges are generally undirected and represent a similarity measure between the two vertices. Labeled instances transmit information through the graph with the goal of labeling unlabeled data [42].
Graph-based methods are based on two important assumptions. The first is the manifold embedding of data into a lower dimensional space enabling the graph representation [42]. The second assumption considers the smoothness of the variation of the labels. Edges with high weights are considered to belong to the same label. Heuristic approaches to compute the weight is discussed in Zhu and Goldberg [41] as follows:

- Fully connected graph: All vertices are connected through a weight-distance function. A function is the Euclidean distance $\|x_i - x_j\|$. Weight and distance have a monotonic decreasing relation.
- kNN graph: A set of k neighbors' vertices are defined for each vertex using the Euclidean distance. For two vertices x_i,x_j, if x_j is part of the k neighbours of x_i, the two vertices are connected through an edge with a constant ($weight = 1$) or the weight can be computed using a distance function. If the two vertices are unconnected, a constant ($weight = 0$) is assigned.
- εNN graph: Two vertices x_i,x_j are connected to each other if $\|x_i - x_j\|<\varepsilon$. The weight will be set to the Euclidean distance if the vertices are connected and will be set to zero if the vertices are unconnected.

Graph based methods have been applied in hyperspectral image classification [46] and natural language understanding [47].

2.4.2. Co-training method

Co-training involves using multiple supervised classifiers [43]. Co-training is based on creating two disjoint subsets of a dataset, i.e., views, and using two classifiers for each view; classifiers contribute to the performance by exchanging their predictions. Two important conditions are assumed in co-training; first, the views are conditionally independent and the second is that each view contains sufficient labeled examples for training the classifiers. Let $X_{unlabeled}$ be an unlabeled dataset and $X_{labeled}$ be a finite labeled dataset $X_{labeled} = [{\chi^1}_l, {\chi^2}_l]$ where ${\chi^1}_l$ and ${\chi^2}_l$ are two independent views on $X_{labeled}$. Two classifiers $F = [f^1, f^2]$ are employed separately on each view and each view is exposed only to one classifier, so f^1 and $f^2$are trained only on ${\chi^1}_l$ and ${\chi^2}_l$ respectively. After creating the training models, $X_{unlabeled}$ is classified using each classifier separately. The most confident label predictions of $X_{unlabeled}$ is exchanged by classifiers and are added to the training data to the other view [41, 45, 48].
Co-training can be used as a wrapper method, which means that any classifiers can be chosen under the Co-training framework. Co-training is used in many applications like e-mail text classification [49], protein subcellular localization [50], classification of images [51] and hyperspectral data from remote sensors [52].

2.4.3. Semi-supervised support vector machine

Semi-supervised support vector machine (S3VM) has been proposed by Vapnik [53] and optimized by Bennett et al. [54]. S3VM was introduced, as an extension to the widely used SVM, to handle the problem of partially labeled data. In traditional SVM, decision boundaries are set to maximize the separation between $X_{labeled}$ whereas S3VM extends the setting to maximize the separation between $X_{unlabeled}$. S3VM is a non-convex np-hard optimization problem that uses the cluster assumption to find the optimal separation employing both $X_{labeled}$ and $X_{unlabeled}$ through using additional constrains in the optimization function of SVM[55]. Different ranges of implementation and optimization have been proposed to solve S3VM [56] in gene analysis [57], text classification [58], in addition, to spectral images analysis [59, 60].

2.4.4. Probabilistic generative models

The idea of this approach is to construct a classifier based on likelihood maximization using both labeled and unlabeled instances. A probabilistic generative model assumes that data is generated from mixture models, which are divided into distinct classes. Both $X_{labeled}$ and $X_{unlabeled}$ are used to estimate the optimal parameters to maximize the probability of the model [61].
If $D = \{(X_{labeled}, Y_{labeled}), X_{unabeled}\}$ is the training dataset, the maximum log likelihood function $\log p(D|\theta_{mod})$ is given by

$$\log p(D|\theta_{mod}) = \sum_{i=1}^{l} \log p\,(y_i|\theta_{mod})\; p(x_i|y_i, \theta_{mod}) + \sum_{i=l+1}^{l+u} \log p\,(x_i|\theta_{mod}) \qquad (2.1)$$

The sizes of $X_{labeled}$ and $X_{unlabeled}$ are given by l and u respectively whereas the parameters of the generative model are given by θ_{mod}.

Eq. (2.1) can be divided into two terms. The first summation represents the log likelihood of supervised learning using $X_{labeled}$. The second term is related to SSL where $p(x|\theta)$ is referred as the marginal probability. The marginal probability estimates the probability of getting $X_{unlabeled}$ regardless of the label [41].

The problem of maximizing the log likelihood $\log p(D|\theta_{mod})$ using hidden data has been discussed by Dempster, et al. [62] under the term 'Expectation Maximization (EM)' in 1977. EM optimization is an iterative method to optimize θ_{mod} resulting in maximizing $\log p(D|\theta_{mod})$ and is carried in two steps [62]. The first step is the expectation step (e-step) where the algorithm generates 'soft labels' of $X_{unabeled}$ given the current model parameters θ_{mod}. The second step is the maximization step (m-step) in which, based on the e-step, the optimal parameters that maximize the log likelihood are found. EM assumes that the prior information, $p(x)$ and $p(x|y)$, of the mixture models are accurate [41], nevertheless, since the labels are missing or limited, the correctness of the model parameters cannot be assessed correctly. To alleviate this weakness, generative models can be applied only for domain knowledge or specialized fields [41] tasks. Another way is to introduce a low weight variable that is associated with $X_{unlabeled}$, so the role of unlabeled data is de-emphasized [41].

Mixture models have been applied in various context like text [37, 63-65] and image classification [66, 67].

2.4.5. Self-training method

In self-training methods, a classifier uses its own prediction to update its training model. In the self-training scenario, the classifiers build a training model based on X_{labeled} using supervised learning. Later, on the learning phase, a new subset of instances $S_i \in X_{unlabeled}$ is selected to predict their labels, where $i \in n$ is the number of subsets. Then the subset S is removed from $X_{unlabeled}$ and added together with their predicted labels to the training dataset $X_{labeled}$. Finally, the classifier is re-trained using $X_{labeled}$ and the labeled subset S_i.This process is repeated until the whole $X_{unlabeled}$ is exhausted or no confident predictions can be further added to the training set.

Usually, in the self-learning process, the subset S contains a few numbers of unlabeled instances. However, the complete set of $X_{labeled}$ and $X_{unlabeled}$ can also be used in the learning process, here the predicted labels might differ between iterations. Self-training is used as a wrapper method, so the prediction function is not restricted to specific classifiers and any classifier can be wrapped in the self-training scenario. On the other hand, self-learning classifiers are sensitive to mislabeling; a wrong prediction can boost itself effecting the retrained model and the overall performance [41]. A vital element in the self-training method is the confidence measure used to select which $x_j \in S_i$ is to be added to the training set. Only the most confident label predictions are added to the training dataset and used to update the training model [37, 41].

Among other applications, self-learning has been applied in human gestures recognition [68], traffic sign classifications [69, 70] and speech recognition [71]. In this thesis, semi-supervised self-training method is used in all the proposed classifiers.

2.5. CONFIDENCE BANDS

Confidence bands are an uncertainty measure of an estimate obtained from limited data, and it defines the area where the true model lies with a pre-defined probability [72]. The certainty predictions in SSL can be employed by introducing confidence bands, which are used to reject possible outliers, i.e., do not lie in the confidence band threshold [73]. Therefore, samples that lie within the confidence threshold are added to the training set, and then, retraining of the classifier is performed using the added data [73]. Confidence bands can be calculated in several ways, for instance, using Monte Carlo [74] or bootstrapping [75]. The confidence band $\sigma_{conf}(\vec{g})$ of the classifier output $\vec{g}$ for a test sample x is measured by

$$\sigma_{conf}(\vec{g}) = \beta\sqrt{g^T(J^TJ)^{-1}g}\sqrt{\sum_{i}^{N} r_i{}^2/v} \tag{2.2}$$

where $\beta = t_{cdf}^{-1}(1 - {}^{\alpha}/_{2}, v)$ is inverse cumulative t-student distribution, α is the probability of the chosen confidence band, we use α =0.05 for 95% confidence bands, and v is the number of degrees of freedom associated with the t-student distribution. The term $(J^TJ)^{-1}$ represents the covariance matrix computed by finding the weighted Jacobian $J = \frac{J_{ij}}{\sigma_i}$ where $J_{ij} = \frac{\partial r_i}{\partial P_j}$ and σ_i are the associated uncertainty of the sample label that may result from a human or self-training. The residual r is the difference between the predicted value and the real value of sample i, and P_j are the classifiers parameters to be optimized [76]. Confidence bands were used in the field of SSL to add certainty to the predictions in gesture recognition [68] and image classification [73]. In this work, we use the output of the proposed classifiers to compute the confidence bands following the procedure presented in [76-78].

2.6. NOVELTY DETECTION

Novelty detection (ND) is a technique used to recognize new test samples, which are unknown to the training model. Depending on the domain of application, the terms one-class classification, outlier or anomaly detection are interchangeably used to refer to novelty detection systems that try to distinguish normal or target samples from abnormal non-target samples. The concept normal/target and abnormal/non-target samples are used to differentiate the known categories or classes, on which the classification model is trained on, from uncommon new data that appears during the test phase. Due to the complexity of real-world systems, it is sometime inapplicable to define a list of all categories that might appear in the test phase. Consequently, conventional classification algorithms are inappropriate for this issue because they will assign a wrong label to the new data sample by employing the predefined categories [38, 79].

Novelty detection is particularly beneficial when a class is extremely under-sampled or when a class is unavailable during the training time. In the first situation, the normal class has few examples to be added to the training dataset; for instance, a particular category happens rarely, so the classification system does not have enough instances to represent this category. In this case, it is better to consider the rare category as novel or abnormal and test it against the more frequently accruing classes. The second situation occurs when the training list is incomplete. Although enough instances are available to form a training model, it is expected that new classes will appear in the future or during the test phase [39]. Therefore, it is important to introduce ND algorithms that can be used to identify new classes which are not yet included in the training dataset.
According to Moya and Hush [79], the one-class classifier is able to identify new training instances (target patterns) and distinguish them from non-target patterns. Obviously, the only available data to the classifier is the class of interest and ND has to distinguish them from all other non-target data [79, 80]. Consequently, a one-side novelty boundary is created based solely on the target class since no other classes are available. On the other hand, in multi-classes classification, data from multiple classes are accessible and the boundary is created depending on data instances from all classes [80].
Following Pimentel et al. (2014) [38], ND is categorized into the following five approaches[38]:

2.6.1. Probabilistic methods

Probabilistic methods are based on using the density estimation of the data to distinguish novel from non-novel instances. The training data set is used to estimate a generative probability density function (pdf) which resembles a model of normality. Using a threshold on the pdf, a test sample can be tested against novelty. This method is similar to the method in Section 2.4.4, where a novel sample is assumed to reside in low dense areas and a known sample is expected to belong to high dense areas [38]. Probabilistic methods can be further divided into:

- Parametric approaches: It is presumed that the normal data is generated from pre-known distributions with pre-calculated parameters. These parameters are finite based on the initial training data and used to fit the model. These distributions can be modeled as a simple Gaussian distribution, as mixture of Gaussian models or a mixture of different distributions, e.g. Poisson or gamma distributions [38, 81]. Common techniques in this category are mixture models [82, 83], extreme value theory [84-86] and state space model [87, 88].
- Non-parametric approaches: In non-parametric approaches, no statistical information about the distribution of the data is assumed. The density function is built using infinite parameters that can grow in size to adapt to the complexity and the form of the data distribution [81]. The main techniques of non-parametric approaches are the kernel-density estimation (KDE) [89, 90] and negative selection [91-93]. Parzen window estimator [94] is a common KDE approach in which the density function is calculated as linear combination of the neighbor kernels at each point in the dataset. Parzen window estimator will be further discussed in Chapter 7.

2.6.2. Distance-based methods

Distance-based methods learn a distance metric to identify the similarity between different samples. They use the assumption that similar data are located near each other, while novel instances are located away from known data. A common technique is the nearest neighbor-based approaches, where the distance, typically the Euclidean distance, is estimated between a point x and the k-nearest neighbor. If the distance is above a threshold, point x is considered novel [95, 96]. Another technique is the cluster-based approach, where the distance between a cluster center and data points that belong to a cluster k is estimated [84, 97]. A point belongs to the known classes, if the distance is within a specific threshold [85]. Though distance based-measures are flexible and do not rely on the distribution of data, these methods depends extremely on the chosen similarity measure, number of neighbors and cluster widths [38].

2.6.3. Domain-based methods

Domain-based methods describe the boundaries enclose the training data. These methods ignore the class density or the sampling procedure and define a domain where the normal data resides [38]. One-class support vector machines (SVM) is one of the most known domain-based methods. A separating novelty boundary is defined as a hyperplane using the closest training points in a mapped space rather than the whole training set [98].
Another variation of SVM is support vector data description (SVDD). SVDD defines a hypersphere with minimum volume that contains all the known training data. This hypersphere comprises the novelty boundary. A test sample is considered abnormal if it lies outside the hypersphere boundary [80, 99].

2.6.4. Information-theoretic techniques

Information-theoretic techniques use uncertainty metrics to derive information contained in the dataset. These techniques presume that abnormal data changes the information related to the content of normal data. In general, metrics such as entropy and Kolmogorov complexity are applied on the whole dataset. If a subset of instances whose removal causes a significant difference in the metric, the subset is considered novel. Information-theoretic techniques do not use the density or distribution of the data but depend extremely on the chosen information theoretic measure [38].

2.6.5. Reconstruction-based techniques

Using the reconstruction-based techniques, test data instances are mapped using a model based on the training set. The objective is to find a mapping that minimizes the reconstruction error. The reconstruction error is defined as a novelty score which is created based on the distance between the test sample and the regression target. Data instances with large reconstruction error are considered novel samples [100]. Reconstruction-based techniques are divided [38] into:

- Neural network-based approaches: Neural networks (NN) one of the most used approaches in ND [101]. NNs are flexible systems that are able to find the

association between the input and the output samples and also capable of detecting novel data [102]. In general, NNs are organized as a series of neurons grouped in layers. These neurons are connected through weighted links. Several architectures and applications of NN have been proposed. The main NN types are Multi-layer perceptron (MLP), Radial basis function (RBF), Auto-associative networks (AANN) and Self-organizing networks (SOM) [38] .

- Subspace approaches: Subspace method use attributes that represent the variability of the data to find an embedding of the training data. This embedding is assumed to be able to map the data into a lower dimensional subspace where normal and abnormal data can be distinguished [103]. Subspace methods are also termed spectral methods. A common technique is Principal Component Analysis (PCA). PCA projects data into a lower dimensional subspace using orthogonal projection. The linear projection is tested for every data instance against the principal component. Instances belonging to known classes comply with the correlation structure of the training data and have a low projection value. On the other hand, instances belong to unknown classes do not satisfy the correlation structure of the training data will have a large projection value [38, 103]. Kernel PCA is an extension of PCA that employs nonlinear projection. Kernel PCA, uses the kernel methods to map the original features into a higher dimensional space and then use PCA to project into a lower dimensional space [104]. These methods are employed in hand-writing recognition [105], breast-cancer detection [106], network intrusion detection [107] and detection of abnormal events in spacecraft components [108, 109].

2.7. THRESHOLD COMPUTATION

A novelty threshold is essential in detecting novel data. Thresholds can be computed using the cross-validation method or using a separate validation dataset. In cross validation, the training dataset is divided into K folds where a fold $k = \{1 \dots K\}$. The training model is built using $K - 1$ folds and the k^{th} fold is hold out and used to validate the training model. This method is repeated until all folds are used as a validation [34]. Validation dataset is a separate labeled training dataset that is used to derive the optimal parameters of the training model. The novelty threshold for each classifier and each class is computed by finding the threshold with the minimum error on a validation dataset. ROC and AUC are used to find the optimal threshold. Brute-force search is applied on all possible values of thresholds per class [110]. The threshold with the minimum false positive and minimum false negative rate is selected as the optimal threshold. During the classification process, when classifying a data point, the threshold is compared to the output of the corresponding classifier. This output or novelty score takes the form of a score or a measure that determines the class membership of the test sample. Scores represent the degree of normality or novelty of a data sample. If the score does not comply the pre-computed threshold, the data sample will be classified as novel [111].

3. Nuclear Magnetic Resonance

3.1. Nuclear Magnetic Resonance (NMR).....18
3.2. One-Dimensional Nuclear Magnetic Resonance.....18
3.3. Two-Dimensional NMR spectroscopy (2D NMR).....20

3.1. NUCLEAR MAGNETIC RESONANCE (NMR)

NMR is an analytical technique used for qualitative and quantitative analyses in numerous applications. Due to NMR's reproducibility, quantitative and non-destructive properties, NMR is considered one of the main instruments used in metabolic profiling. The analysis of metabolites allow a differentiated prediction of the health status and potential health risks of a patient [112].
NMR is used in multi-component mixture analysis of biological fluids such as plasma, urine, and serum in addition to tissues. The primary goal of metabolic profiling using NMR is the prediction, diagnosis, monitoring and prognosis of diseases as well as optimizing medication efficacy [113]. In general, all metabolites have a known and reproducible NMR pattern. Using these patterns, NMR can be used to investigate the metabolic composition of complex biological samples. In this chapter an introduction to 1D and 2D NMR is introduced.

3.2. ONE-DIMENSIONAL NUCLEAR MAGNETIC RESONANCE

NMR spectroscopy is based on the existence of the nuclear spin angular momentum inducing a magnetic atomic moment. Once an external magnetic field is applied on a nucleus a splitting into ground and an excited spin states is induced [114]. Radio frequency (RF) is used to promote energy transitions between these states. The. RF frequency required to induce energy transition depends on i) the nucleus type (e.g., ^{1}H or ^{13}C), ii) the chemical environment of the nucleus and iii) when the field is not uniform, the typical nuclei location in the magnetic field [115]. In addition, the distribution of electrons in the chemical bonds effects the local magnetic field [116]. As illustrated in Figure 3.1, after applying a RF pulse, the nuclei transfer to the exited state. After the RF radiation ends, the external magnetic field B_0 acts upon magnetization M of the atomic nuclei, which starts a precession around the z-axis of the external field with a characteristic frequency. The x and y components of the magnetization after the RF pulse is measured with a receiver coil and is called free induction decay (FID). The time domain signal is converted via a Fourier transformation to the frequency domain. Depending on the molecular structure and the chemical environment of the excited nuclei characteristic frequency are seen in the spectrum [116].

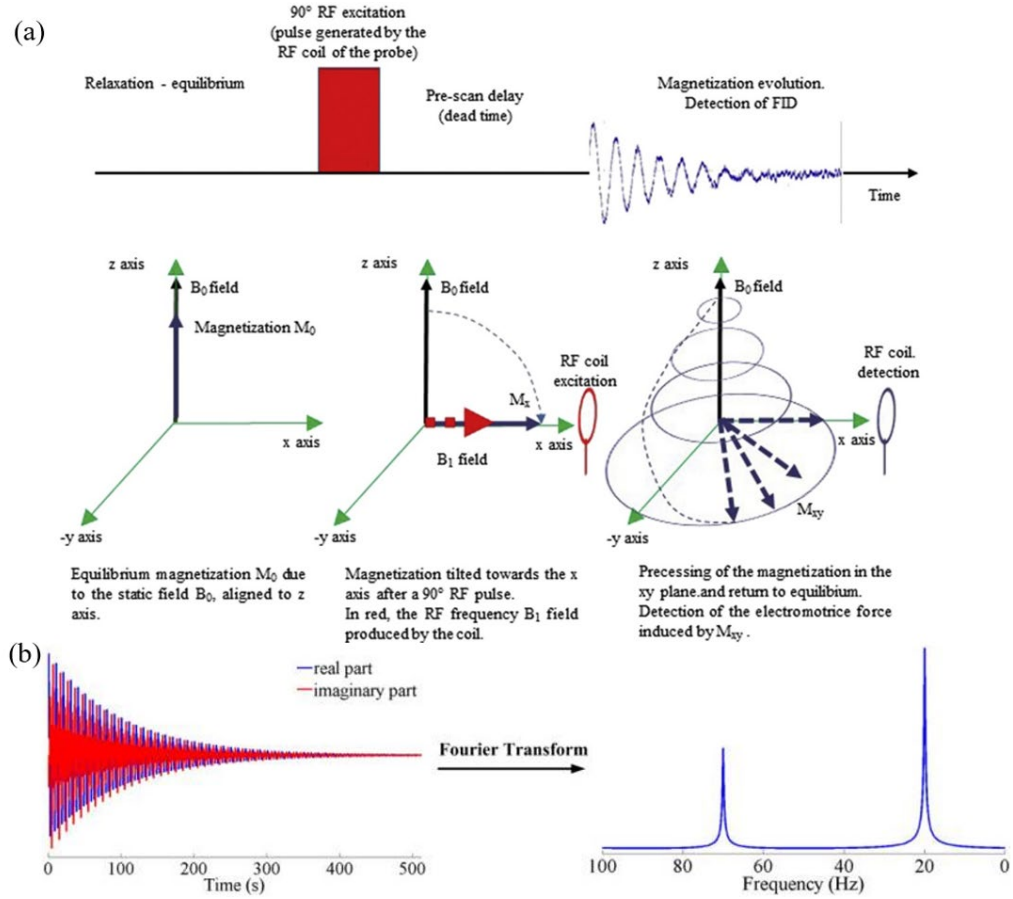


Figure 3.1: One-Dimensional NMR spectroscopy pulse-acquisition and Fourier transform. **(a)** Pulse-acquisition 1D NMR experiment. The magnetization at equilibrium is aligned along the B_0 direction (typically z-axis), ideally at the end of the relaxation delay. After 90° pulse (RF excitation pulse), magnetization is flipped in the x-y plane and then the precession of the flipped magnetization gives the FID which is detected with the NMR detector (typically inductance coil). FID is a form of an NMR signal where the magnetization is flipped by 90° B_0 (conventionally along z) using a 90° pulse leads to non-equilibrium magnetization [117]. **(b)** Time domain FID plotted data and its corresponding 1D NMR spectrum by the Fourier transform (FT). The diagrams in (a) and (b) are customized from the literature [118, 119].

Tetramethylsilane (TMS) in organic solvents or sodium 2,2-dimethyl-2-silapentane-5-sulfonate (DSS) in aqueous solutions, are recommended as universal primary frequency references. The methyl ^{1}H signal chemical shift of TMS is equal to 0 ppm and therefore frequencies of chemicals shifts are calibrated according to the ^{1}H or resonance of TMS [120]. Moreover, 3-(trimethylsilyl) propionic acid sodium salt (TSP) are commonly used for NMR studies is used as reliable internal chemical shift reference of compounds [121-123]. The chemical shift ranges of ^{1}H-NMR of organic compounds are illustrated in Figure 3.2, the range is customized from 0-11 ppm [124].

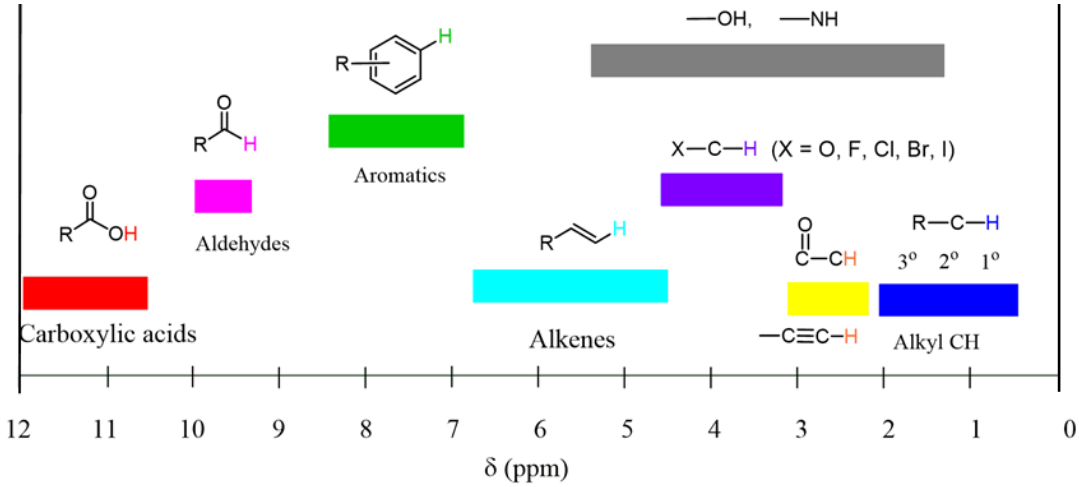


Figure 3.2: Schematic diagram of ^{1}H chemical shift ranges for organic compounds . Adapted from [124].

3.3. TWO-DIMENSIONAL NMR SPECTROSCOPY (2D NMR)

In the previous section, acquisition of 1D NMR has been described. Though 1D NMR can be obtained in a relatively short time, obtaining a good signal to noise ratio requires longer data acquisition times. In addition, due to the short chemical shift range and the limited spectral widths in 1D NMR, there is an increased probability of overlapping spectrum especially in complicated mixture of bioorganic molecules such as the example shown in Figure 3.3 [125]. Dense and overlapped 1D NMR spectra are hard to analyze and prone to wrong annotations. Therefore, introducing multi-dimensional NMR techniques can increase the spectral resolution and alleviate spectral crowding which can credibly detect and identify more metabolites than 1D NMR [11, 112].

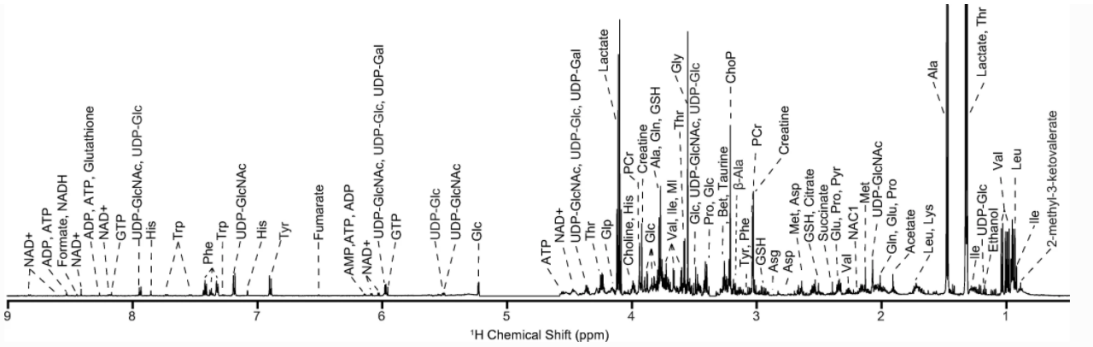


Figure 3.3 : ^{1}H NMR spectrum at 600.13 MHz of a HeLa cell extract showing metabolite annotated on the spectrum. Metabolites abbreviations: betaine, Bet; phosphocholine, ChoP; pyroglutamate, Glp; glutathione, GSH; N-acetyl 1, NACl; myoinositol, MI; phosphocreatine, PCr; pyruvate, Pyr; UDP-galactose, UDP-Gal; UDP-glucose, UDP-Glc; UDP-N-acetyl-glucosamine, UDP-GlcNAc [125].

While 1D NMR provides a correlation between frequency and intensity, 2D NMR correlates two frequencies. The intensity is represented as third dimension and plotted as contour lines of the two frequencies. 2D-NMR spectroscopy is used to provide information about the correlation between nuclei through-bond (J-coupling) or through-space (Nuclear Overhauser effect) to observe the molecular structure in detail. Usually,

in 2D NMR, the direct detection dimension is ^{1}H while the indirect (transient) dimension could be ^{1}H, ^{13}C, ^{15}N, ^{31}P, or other nuclei. 2D NMR power lies in its capability to resolve overlapping peaks [11].

Figure 3.4 shows a simple two-dimensional NMR pulse-sequence. Like 1D, after applying an RF pulse to the nucleus, the system starts to relax back to equilibrium in the z axis, nevertheless, the generated FID is not recorded but is left to evolve for an evolution period t_1 and a transfer of magnetization happens between coupled nuclei. A second RF pulse with a frequency resonating with the second nuclei is applied. The excitation of the coupled nuclei starts decaying and the resulting FID is acquired by the coil. The resulting FID contains information related to the coupled nuclei due to transfer of magnetization. Transfer of magnetization between the coupled nuclei is translated into cross peaks in 2D spectra. A 2D FT is applied on the two FIDs to transfer time domain signals into frequencies [126, 127].

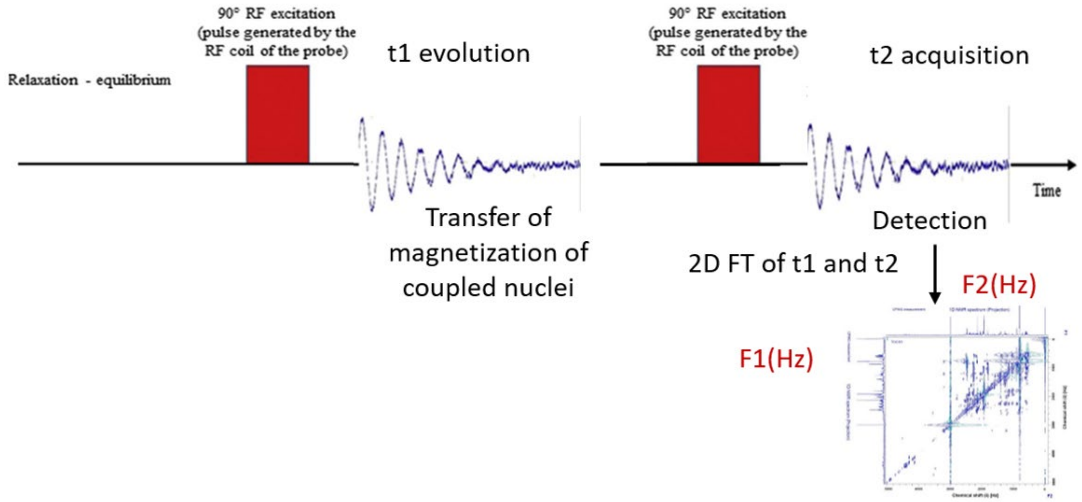


Figure 3.4: Two-Dimensional NMR spectroscopy pulse-acquisition. Adapted from [118]

In 2D NMR spectra, each peak is determined by plotting the horizontal frequency (F2) versus the vertical frequency (F1) and contour lines are used to represent the intensity of the signal [128-130]. The following are some of the most common 2D NMR spectrum types.

3.3.1. Correlation spectroscopy (COSY)

COSY is considered the simplest type of 2D NMR experiments. COSY is a homonuclear experiment which establishes the coupling between two close protons of two hydrogen nuclei (^{1}H) indicating connectivities up to four bonds [131]. The diagonal of COSY spectra resemble the spectrum of the 1D NMR experiment, whereas the couplings between pairs of protons are indicated by cross- peaks on the off-diagonal [11, 26].

The COSY experiment is relatively fast and simple to analyze. Nevertheless, in complex mixtures, short bond range and spectral overlap increases the complexity of the COSY spectrum and harden the analysis process [11].

3.3.2. Heteronuclear single quantum correlation spectroscopy (HSQC)

^{1}H–X HSQC is a heteronuclear 2D NMR spectrum which shows the correlation between two different chemical nuclei, for instance ^{1}H and Carbon-13 (^{13}C), Nitrogen-15 (^{15}N) or Phosphorus-31 (^{31}P) [132-134]. HSQC is used widely in protein NMR where the horizontal axis shows the chemical shifts of protons and its correlation with ^{13}C, ^{15}N or ^{31}P is shown on the vertical axis. HSQC spectrums provides less spectral overlapping and offers a more detailed fingerprints of molecules [134, 135]. Nevertheless, there are strong coupling effects that influence the HSQC experiment especially when the NMR magnetic field is weak [136, 137]. Furthermore, there is the disadvantage of missing spin system information, as all cross-peaks are independent of each other in HSQC [138, 139]. Though HSQC can experience strong peak deviations and loss of intensity, the sensitivity of HSQC is generally inadequate for metabolomics studies [140, 141]

3.3.3. Heteronuclear multiple-bond correlation spectroscopy (HMBC)

HSQC shows the heteronuclear correlations only to directly bounded proton. Therefore, coupled nuclei which are not in direct one-bond relation are not detected. On the other hand, HMBC reveals long-range heteronuclear correlations between protons and a different chemical nucleus which are separated by chemical bonds which range between 2 to 4 bonds. In HMBC, the direct one bond is eliminated through filtering only small j-coupling constants by introducing a longer delay allowing evolution of the two or three-bond. To analyze the whole spin system, a combination of HMBC and HSQC is recommended [11, 142].

3.3.4. Total Correlation Spectroscopy (TOCSY)

Group of spins that are coupled are called a spin system. TOCSY shows the connectivity between pairs of spins and the total spin system [126]. TOCSY correlates between the coupled protons for continuous chains of protons and is not only restricted to three or four chemical bonds like COSY [11, 26]. Therefore, TOCSY reflects the chemical shift information of all members of the spin-system [143]. The efficiency of the TOCSY experiment is related to the magnitude of the J-coupling, the mixing time, and the distance between the coupled nucleuses. The closer the chemical shift distance, the larger the resonance of the spin system. TOCSY spectrum is a homonuclear 2D experiment, which is shown as a symmetrical 2D of two diagonally symmetrical contour plot where the diagonal represents the 1D spectrum, and the cross diagonal represents the correlation between the nuclei. The contours shows the amplitude of a signal as a function of the F1 and F2 frequency axes [126]. In TOCSY, diagonal peaks represents singlet patterns in 1D-NMR experiment and do not indicate any coupling, while the off-diagonal cross peaks correspond to coupled nuclei [126]. Figure 3.5 displays a 2D ^{1}H-^{1}H TOCSY of a urine sample [143].

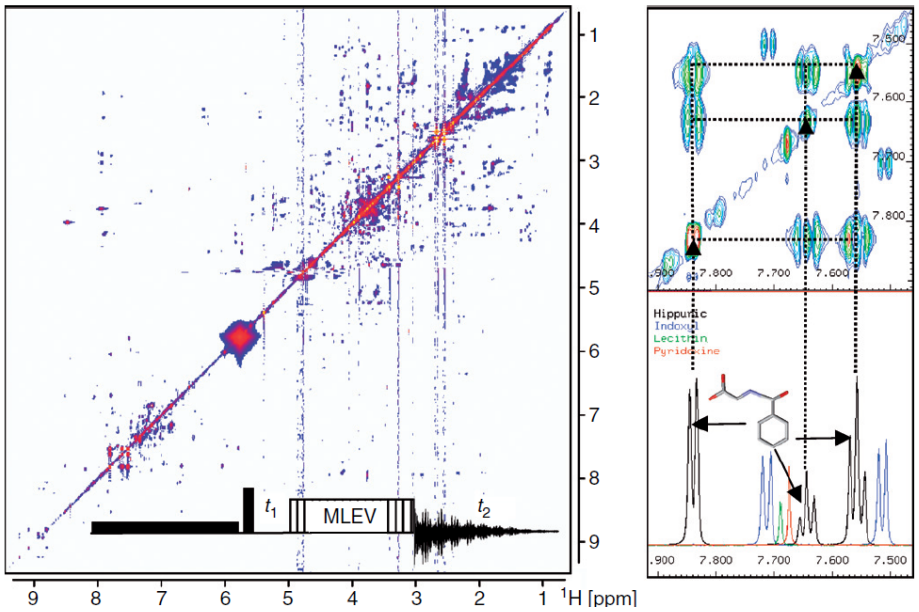


Figure 3.5: 2D ^{1}H-^{1}H TOCSY contour plot of a urine sample (left), the top right insert shows an expanded contour region between 7.5 ppm and 7.9 ppm of correlating protons. The downright inserts show the 1D projection of the corresponding ppm area. Coupled components can be seen in the same colors. Lecithin and pyridoxine are singlet signals and appear only diagonally. The MLEV pulse sequence is used to acquire TOCSY spectra and the MLEV mixing sequence is used to transfer magnetization along J-coupled bonding [143].

The manual spectral deconvolution is dependent on user experience and is a severe bottleneck in the field [144, 145]. Additionally, it is an impractical and tedious process, especially for high-throughput applications and complex biological mixtures [146, 147]. Semi-automated approaches have been developed to decompose TOCSY spectra into individual components matching- an NMR databases for identification [148]. DemixC is a semi-automated technique that deduces 1D cross-sections (row) of a 2D TOCSY spectrum that does not exhibit many peak overlaps [148], and peak fitting is used to extract peak positions from a TOCSY spectrum [148]. Frequently, metabolomics samples are composed of hundreds of individual components, which may result in overlapping peaks and, consequently, problems of the DemixC method [148]. Therefore, the Demixing by Consensus Deconvolution and Clustering (DeCoDeC) is preferable to dealing with mixtures of higher complexity [149]. DeCoDeC identifies peaks apparent in specific pairs of TOCSY 1D cross-sections so that overlapping peaks associated with other metabolites are eliminated [148]. Significant limitations of both approaches arise because of the peak shifts due to matrix effects, which is the common case in metabolic profiling investigation of real-time evolution measurements [146].

4. NMR-Based Metabolomics

4.1. 1D NMR Metabolite Spectra 25
4.2. 2D NMR Metabolite Spectra and Metabolite Assignment.....26
4.3. Automated Metabolite Identification.....27
4.4. Related Work 28

Fundamentally, metabolites are the input and the output of any biological process, and their associated biomarkers are linked to a broad scope of disease, disorders, genetic reformation, and environmental settings. Therefore, metabolic studies are suitable approaches for research in epidemiology, cancer research, biotechnology, drug design and toxicology [13].

The first NMR measurement was reported in 1938 by Isidor Rabi who has been awarded a Nobel Prize for his work. However, the potential capabilities of NMR to study living cells were not investigated until early 1980. Since 1985, NMR has been used to measure biological tissues and fluids [112]. NMR is established as one of the principal tools for metabolic studies and multi-component mixtures analysis for the following reasons: NMR offers detailed chemical information of metabolites in a short period of time [11, 150]. Second, NMR is a non-destructive and a non-invasive technique, consequently, NMR can be used for living cells and real-time metabolic analysis of the same sample without damaging it [11, 150]. Third, the results of NMR are highly reproducible; if biological samples are stored below -80 °C, these samples can be recovered and repeatedly re-measured to give the same recurrent results. So, researchers test a sample using NMR for initial evaluation, store it, and then re-measure the same sample using NMR for further analysis [11]. Fourth, even for NMR measurements of highly complex biological mixtures, sample preparation in NMR requires minimal or no sample preparation before moving the sample to the NMR instrumentation [3, 11, 150, 151]. Moreover, NMR can be applied for in-vitro and in-vivo metabolic profiling. NMR analysis verifies the possibility of translating the finding of in-vitro experiments to in-vivo medical applications [11, 152]. Accordingly, NMR practices and evolutions have continued to emerge and to expand [153]. Figure 4.1 shows the growth of number of publications related to identifying metabolites in biological systems using NMR in the past years. Nevertheless, several limitations are associated with analyzing complex biofluids using NMR, such as low resolution and sensitivity [154].

4.1. 1D NMR METABOLITE SPECTRA

Most NMR metabolic profiling studies employ statistical pattern recognition, such as partial least-squares discriminate analysis or PCA methods on 1D 1H NMR spectra. However, because of the large variability of molecular concentrations in living systems, statistical analysis of 1D NMR is biased toward distinguishing fluctuations in the low

concentration metabolites. These metabolites are usually hidden due to spectral crowding [155].

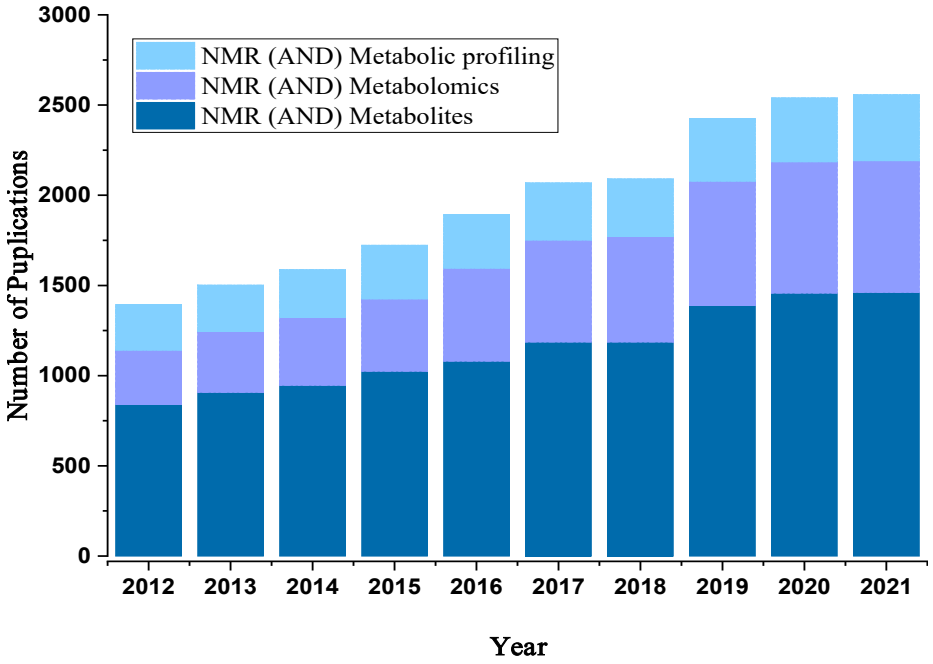


Figure 4.1: A summary of the number of publications from years 2012 to 2021 on NMR in metabolomics. This figure is generated using topic search on Web of knowledge applying the keywords: 'NMR (AND) Metabolites', 'NMR (AND) Metabolomics' and 'NMR (AND) Metabolic profiling'.

This issue is further complicated by small but critical chemical shifts due to fluctuation in pH values, ionic strength and other factors summarized as matrix effect [13]. Matrix effect results from the variation between the responses of a component in a standard solution and its response in biological matrix. Matrix effects are hard to predict and are effected by interfering components such as lipids and protein [156]. Two dimensional NMR is used to overcome the limitation of 1D NMR, to increase spectral resolution and dispersion which helps in determining overlapping metabolites [11, 142, 157].

4.2. 2D NMR METABOLITE SPECTRA AND METABOLITE ASSIGNMENT

Metabolite identifying using 2D NMR techniques can reveal more information about the studied mixture. All 2D NMR spectra use the principle of adding a second dimension by recording a sequence of 1D NMR spectra incorporating a series of time intervals. A Fourier transformation is applied to these time intervals to generate an orthogonal second frequency domain [158].

2D NMR increases the signal dispersion and displays connectivity and chemical bond information. In addition, homonuclear NMR techniques such as ^{1}H-^{1}H COSY and ^{1}H-^{1}H TOCSY and heteronuclear techniques such as ^{1}H-^{13}C/^{15}N HSQC and ^{1}H-^{13}C/^{15}N HMBC describe direct and indirect coupling and correlation between ^{1}H-protons and a second nucleus such as ^{13}C-carbon or ^{15}N-nitrogen. Moreover, multiple techniques can be combined to reveal more information on biological mixtures. For instance, measuring 2D HSQC, TOCSY, and HSQC–TOCSY subsequently allows a comprehensive analysis of samples [153, 158-160].

Attempts to overcome signals overlapping in 1D NMR spectra of metabolites have included (i) skyline projection of 2D J-resolved spectroscopy to obtain a broadband proton-decoupled 1D spectrum [161], (ii) use of isotopically enriched metabolites [162] and (iii) sample fractionation [163].

A 2D ^{1}H-^{1}H TOCSY experiment with the zero-quantum filter technique developed by Thrippleton et al. [164] to obtain phase peaks can be used to obtain a metabolic profile of a biological sample and at the same time perform proper assignment for the metabolites in spite of crowding 1D spectra [164]. The experiment allowed more accurate quantitation of low-abundance metabolites. This approach applied on 1D proton and 2D TOCSY NMR was employed to analyze the metabolic profiles of urine obtained from wild-type and Abcc6-knockout mice.

4.3. AUTOMATED METABOLITE IDENTIFICATION

Metabolic profiling encompasses the investigation of metabolites concentrations, systematic metabolic variation that are caused by different drugs, dieting, microbiological causes, gene modulation or new stimuli for the purpose of the characterization of the effects these interactions [165]. Due to the nature of the biological fluids, cells and tissues, metabolites are changing to reach a dynamic equilibrium in the body. As a result, any abnormal biological process, will cause a metabolic deviation in the body and biofluids which can be related to the diagnosis or prognosis of these abnormal biological processes [150, 166]. Detecting abnormal perturbations can reveal specific diseases or therapeutic status. NMR spectroscopy is one of the most powerful tools that are used in the multicomponent analysis of biofluids such as urine, blood plasma or tissue abstracts [166].

Several NMR related limitations originate from its limited sensitivity and resolution [167]. Though major efforts to lessen these limitation, the complexity of biological mixtures demands further developments and enhancements for detection, identification, and quantitation of complex biological mixtures [3]. One of the major challenges of NMR spectroscopy is peaks overlapping. Peaks overlap and chemical shift is expected to occur not only between different molecules, but also within the same molecule in the case of complex multiplets overlap. Though the identification of metabolites in 2D NMR spectra is relatively simpler than 1D NMR [166], the straightforward identification of metabolites in 2D NMR is valid only to first orders systems with weak coupling. In 2D NMR, the identification of metabolites which appears on a relatively low intensity, or that have peaks which are partially or totally overlapped is a complicated task [166]. Consequently, in complex experimental measurements, new peak shifts, misaligned peaks as well as peaks with slight deviation of the expected peak shape make metabolic profiling of NMR measurements a challenging task [13, 167]. Therefore, even for experts and researchers, the manual analysis of NMR spectra is an elaborative, complicated and a time-consuming task. In addition, the manual analysis of 2D NMR spectra is prone to error and miss-assignment in cases of complex mixtures with several overlapped metabolites and in high-throughput applications [168]. In general, the classical manual analysis of 2D NMR

spectra can be a bottleneck in the research and experimental workflow in NMR spectroscopy.
Consequently, automating the process of metabolic profiling in biological mixtures will support and speed the process of 2D NMR analysis significantly. Moreover, creating an automatic assignment system will enhance the process of knowledge transfer, so even nonexperience researchers will be able to analyze and assign the metabolites that appears in the 2D NMR spectra [168, 169]. NMR spectroscopy and Machine learning (ML) create a promising interdisciplinary research area that will achieve a notable progress in NMR spectroscopy leading to an advancement of the diagnostic and prognostic use of biomarkers in addition to drug design and discovery [170].

4.4. RELATED WORK

Introducing ML to serve as an analysis tool for NMR appears to be a reasonable effort. MetaboAnalyst 3.0 is an R-based tool for metabolomics studies (www.metaboanalyst.ca) in 1D NMR. MetaboAnalyst 3.0 [171] enables metabolomics analysis, visualization and interpretation [171] using the metabolome libraries HMDB [172], KEGG [173] and SMPDB [174]. In addition, MetaboAnalyst 3.0 has been enhanced by a biomarker analysis module for biomarker identification and features ranking using PCA clustering, partial least squares - discriminant analysis PLS-DA classification, t-tests and ANOVA [171]. For raw spectral data processing, MetaboAnalyst 3.0 users have to use an external software for the simplification and processing of the spectrum before using the tool [171].
Another R-based tool for analyzing 1D spectra is BATMAN. 1D NMR spectroscopy is commonly used for estimating concentrations of chemical substances in solution. BATMAN metabolic spectral resonance patterns are derived from the metabolites library (HMDB) [172] by incorporating this information into a Bayesian model, which deconvolve NMR spectral resonance peaks to identify metabolites and to measure their concentrations. The reference spectra are stored in the form of chemical shifts, J-couplings and multiplet intensity ratios [13]. These properties are used in the sense of a prior probability in a Bayesian framework, allowing for slight deviations of the observed spectral parameters from those of the reference spectra due to pH and ionic strength [13].
1D NMR spectroscopy is commonly used for estimating concentrations of chemical substances in solution [12]. However, in complex mixtures of chemical species such as in metabolomics, strong peak overlaps are encountered and then 2D NMR is an alternative approach since peaks superposition in 1D NMR spectra can often be separated in 2D NMR spectra [166].
Several computer implementations have been proposed to enable NMR spectral processing and cross peaks identification of 2D NMR spectra. COLMARm web server is an online available platform that incorporates three 2D NMR spectra for the purpose of simultaneous analysis [175]. COLMARm operates in two stages; first, an HSQC spectrum is uploaded by the user, compared against a unified database from Biological Magnetic Resonance Data Bank (BMRB) [176] and The Human Metabolome Database (HMAB) [172] and a matched list of metabolites is created. On the next step, the matched list is validated against the correspondent TOCSY and/or HSQC-TOCSY spectrum. This

method uses pattern matching with referencing points. These referencing points could be standard referencing or commonly appearing metabolites. COLMARm needs human interventions in the validation step, so this method is not considered fully automatic.

Another category of metabolite identification is structure elucidation and identification [177]. NMR is one the of the most established procedure in this category [151]. The term structure elucidation is defined as is the procedure of identifying the chemical structure of a molecule via the determination of the chemical elements numbers and types which constitute the molecule [178].

Sheen et al. [179] describe a procedure for spectral outlier classification in 2D NMR using protein chemical structure imported from the NISTmAb International Multilaboratory NMR experiment [180].This method incorporates symmetric Kullback-Leibler divergence as a similarity measure between spectra. A similarity score based on each spectrum and other similar spectrum is calculated. If the similarity score excessed a confidence limit, a spectrum is considered as outlier [179].

A Bayesian framework has been used for the problem of the assignment of peaks in 2D NMR spectra in different formulations. In [181], 2D NMR spectra are modeled as a mixture of bivariate Gaussian densities. To estimate the positions of the peaks, the adaptive Markov chain Monte Carlo (MCMC) algorithm is used. A list of candidate peaks of the highest amplitude is created and the posterior probability of each candidate peak is calculated [181]. Another technique that use the Bayesian framework and Pictorial Structures is proposed in [182]. It is assumed that metabolites can be represented as vectors of chemical shift $z \in M$ and a spectrum can be represented as a set of spectral images $\mathrm{I} = \{I^1 \dots I^k\}$. The assignment problem is modeled as calculating the maximum posteriori estimation (MAP) of z by $z_{MAP} = \arg\max_z p(z|\mathrm{I})$. The spectral image likelihood $p(z|\mathrm{I})$ can be estimated using Bayes' theorem as $p(z|\mathrm{I}) \propto p(\mathrm{I}|\mathrm{z})p(\mathrm{z})$ [182]. A more recent approach using NMR spectral line shape in 2D J-resolved NMR is presented in [183]. The NMR Lorentzian distribution and the associated parameters like B-spline tight wavelet frames and theoretical templates are incorporated into the Bayesian method. Online databases are used to create an estimate of prior distributions of NMR related parameters like J-coupling constants, peak shape parameter, multiplet chemical shift and global peak width. Markov Chain Monte Carlo estimate is used to perform the posterior inference based on the likelihood and prior functions. This approach is related to 1D NMR analysis through BATMAN tool mentioned previously [13, 184]. Another peak assignment approach [185] which incorporates the shape of the peak on the 2D spectrum is introduced in [185]. After selecting peaks that are within a predefined threshold, a technique called the Histogram of Oriented Gradients (HOG) is used to extract the features of the peaks. HOG transfers the image of the peak from the 2D spectrum into a matrix of features through shape mapping. These features are trained and tested using SVM classifier [185].

Neural networks have been exploited in NMR for the reconstruction, denoising of spectra, chemical shift prediction and automatic peak picking [145]. Mostly, these applications are implemented using mainstream libraries like Tensorflow [186] or Matlab Deep Learning Toolbox [187, 188]. For chemical shift prediction, multiple types of

features have been used as feature space for the training dataset. The first and most common is the structure of molecule where the relationship between the chemical shift and the environment and structure related information of the compound is estimated. In [189] a peak list is created from different 2D spectra, such as ^{1}H-^{15}N HSQC, ^{1}H-^{13}C HSQC, HCCH-TOCSY, 15N-edited NOESY and ^{13}C edited NOESY. These peaks are manually inspected by NMR analysis tool, KUJIRA [190] and converted to grayscale 2D and 3D images. The images are used to build a Cognitive Neural Network Tool Kit from Microsoft [191] for the purpose of automatic peak identification. These peaks are then provided to the tool with CYANA [192] for signal assignment and structure elucidation [189]. SMART and SMART 2.0 [193, 194] are based on training a deep convolutional neural network (CNN) of Siamese architecture [195] to asset the uniqueness of the compounds, in addition to the annotation of known compounds in biological mixture. SMART 2.0 is trained on 25434 HSQC spectra from the JEOL database [194]. This tool is designed to facilitate the structural elucidation of known compounds and discover new categories through using the CNN to create clusters by incorporating PCA and performing the annotation based on similarity metric [193, 194].

Several studies described the metabolism of MSCs and metabolic changes due to adipogenic [196], osteogenic [197, 198], and chondrogenic differentiation [199]. Stem cell osteogenic differentiation of hMSCs for 21 days based on 1D NMR has recently been studied [200, 201]. They mainly considered the lipidomic and amino acid characterization of osteogenic stem cells using PCA and partial least squares discriminant analysis. Human embryonic stem cells were studied to monitor the intracellular and extracellular metabolic dynamics through directed and non-directed differentiation using 1D NMR. Similarly, PCA, least square analysis and ANOVA test were used to compare the differentiated and undifferentiated cells [202, 203].

5. Datasets

5.1. NMR data acquisition and processing 32
5.2. TOCSY crosspeak picking and de-noising 32
5.3. Breast cancer tissue cells.....33
5.4. Adipose tissue-derived human Mesenchymal Stem cells 36
5.5. Data Representation 37

5.1. NMR DATA ACQUISITION AND PROCESSING

^{1}H NMR measurements were performed using HR MAS ^{1}H NMR probe head operated by a Bruker Avance III 600 spectrometer at 600.13 MHz for ^{1}H at 276 K. HR MAS spinning frequency was set to 5 kHz, and the magic angle was adjusted typically according to the KBr measurement. The B_0 magnetic field shimming was performed manually until the linewidth of the alanine signal at 1.46 ppm was adjusted to fall within the range of 1.20–1.83 Hz. Metabolites were deduced from the ^{1}H NMR spectrum based on expert knowledge with the assist of ^{1}H-^{1}H TOCSY, ^{13}C-^{1}H HSQS and the Chenomx NMR Analysis Software from Chenomx Inc. Details are presented in [204-206]
To avoid blurring of multiplet pattern, ^{1}H-^{1}H TOCSY was recorded with suppressed zero-quantum coherences [164]. TOCSY were measured with a spectral range (SWH) of 7 kHz in both F2 and F1 dimensions. Mixing time and relaxation delay were set to 80 ms and 1 s, respectively. Zero filling was performed to 16K and 128 data points in F2 and F1 dimensions before 2D Fourier transformation [204-206]. The spectral widths in the F2 and F1 dimensions can be adjusted or enlarged according to the area of interest in the TOCSY. The 1D NMR spectral projections on the F1 and F2 axes are external projections from extra 1D NMR measurement using the CPMG pulse sequence with embedded water suppression by excitation sculpting. CPMG was used to suppress protein, lipids and other macromolecules and it was recorded employing 400 echoes with 1 ms echo time.

5.2. TOCSY CROSSPEAK PICKING AND DE-NOISING

The cross-peaks entries in F2 and F1 dimensions in ppm and Hz are deduced from the 2D contour lines of the experimental 2D TOCSY NMR spectrum by employing the automatic peak picking function (pp2d) in TopSpin 3.6 provided by Bruker for acquisition and processing. Before applying automatic peak picking, the contour projection magnitude threshold was adjusted for every ppm range in F2 dimension according to the amplitude of the 1D NMR spectrum internal projection on F2 axis to avoid picking

artifacts and noise cross peaks. Afterward, the collected peaks were listed and transferred as text file for data de-noising and artifact cross-peak elimination. in TOCSY spectrum, every real cross-peak appeared in the upper diagonal (F2, F1) due to the J-coupling should have a mirror (transpose) cross-peak in the lower diagonal (F1, F2) within tolerance threshold of ~30 Hz, based on that we could exclude cross-peaks that do not fulfill this criterion. Moreover, most cross peaks in vicinity of water and solvent signals are associated with T1-noise [28]. Fortunately, T1-noise appears in TOCSY spectrum as random or semi-random spurious streaks along the indirect F1 dimension of a 2D NMR spectrum and they have no transpose (mirror) in the lower diagonal entries (F1, F2). Typically, no metabolite signals in vicinity are taken for assignment, since other characteristic peaks in different F2 and F1 ranges can be considered. It is worth mentioning that metabolites that have no coupled protons will show singlet signals in 1D NMR and therefore, no cross-peaks in TOCSY. Such signals will only have contour projections in the diagonal. Typically, 2D TOCSY spectra provide information about correlated protons of the same spin system. However, peaks in the diagonal can be used as a part of the data to solve the issue of metabolites with no intrinsic coupling if they are not severely overlapping. A spectroscopic more favorable approach would be correlation measurements between ^{1}H-^{13}C HSQC [207, 208]. The term 'targeted metabolic profiling' is used for the analysis of certain molecules or functional groups rather than the whole spectrum, on the other hand, in this work, non-targeted metabolic profiling of the whole spectrum is used. Non-targeted metabolic profiling is an all- inclusive and comprehensive analysis of the whole spectrum and all peaks above a predefined intensity threshold are selected and analyzed. Automating non-targeted metabolic profiling has unlimited perspective in overcoming the inherent obstacles in non-targeted 2D NMR analysis [209, 210].

5.3. BREAST CANCER TISSUE CELLS

Breast cancer is considered one of the most frequent tumors and the leading cause of cancer death among women [211, 212]. Although, in its early stages, breast cancer has a curability rate of 70-80%, progressed breast cancer can be mortal [213]. Recent studies target to detect the potential and common metabolic signature for the purpose of early diagnosis, prognosis evaluation and to improve the realization of the metabolic pathobiology of breast cancer.

The breast cancer tissue data used in this work has been previously analyzed and published [204]. The work was part of a comprehensive study focusing on the heterogeneity of cancer tumor tissues. Breast tumor tissue samples from 18 patients were analyzed. After surgery, a specimen for pathological diagnosis was immediately procured and the remaining tissue was snap frozen and stored at –80°C within 10 minutes. Six cores each taken from a different patient were analyzed blindly by HR MAS ^{1}H-NMR [204, 206].

A 1D NMR spectrum of the sample was measured, analyzed, and assigned based on expert knowledge with the help of the Chenomx NMR Analysis Software. A number of 27 metabolites were assigned in the measured real breast cancer tissue sample as

following, namely: 'Valine', 'Isoleucine', 'Leucine', 'Lysine', 'Glutamate', 'Alanine', 'Glutamine', 'Aspartate', Sn-Glycero-3-phosphocholine (GPC), 'Serine', 'O-Phosphoethanolamine', 'Ascorbate', 'Myo-Inositole', 'Lactate', 'Proline', '3-Hydroxybutyrate', 'O-Phosphocholine', 'Threonine', 'Glutathione', 'Inosine', 'Beta-Glucose', 'Alfa-Glucose', 'Tyrosine', 'Phenylalanine', 'Uracil', 'Taurine' and 'Methionine'. Figure 5.1 shows the ^{1}H-^{1}H TOCSY spectrum of a breast cancer tissue sample studied in this work at 600.13 MHz with mixing times (τ_m) of 80 ms. The 2D TOCSY spectra were recorded using a pulse sequence that suppresses zero-quantum coherences [164] to avoid blurring the multiplet patterns with a relaxation delay of 1 s. In this way, the resulting multiplets exhibit the same structure as in 1D NMR spectra, which facilitates classification. Measurements with a high indirect frequency resolution can only be obtained by a subdivision into many time increments, resulting in long measurement cycles. The spectral range was set to 7 kHz in both dimensions, 16K and 128 data points acquired in the horizontal and the vertical dimension (F2, F1), respectively. Before 2D Fourier Transform, zero fillings were performed to 32K and 1K data points in the horizontal and vertical dimensions, respectively. The spectral widths in the two dimensions were acquired on spectral range of 12.00 ppm to cover all possible metabolites chemical shifts. The spectral ranges up to ~ 9.0 ppm (5600 Hz) in F2 and F1 dimensions was considered since the cross-peaks of the metabolites in the TOCSY spectrum were appeared only in these spectral ranges. The NMR experiment has been acquired at 279 K. The peak (F2, F1 in Hz) entries are deduced from the experimental 2D TOCSY NMR spectrum of the real breast cancer tissue from the 2D contour lines using the automatic peak picking function (pp2d) in Bruker TopSpin 3.6. The peak picking level was adjusted based on the contour projection magnitude threshold to avoid picking artifacts and noise peaks. Peaks are annotated in the TOCSY spectrum using the red square symbol associated with peak number, as illustrated in Figure 5.1.

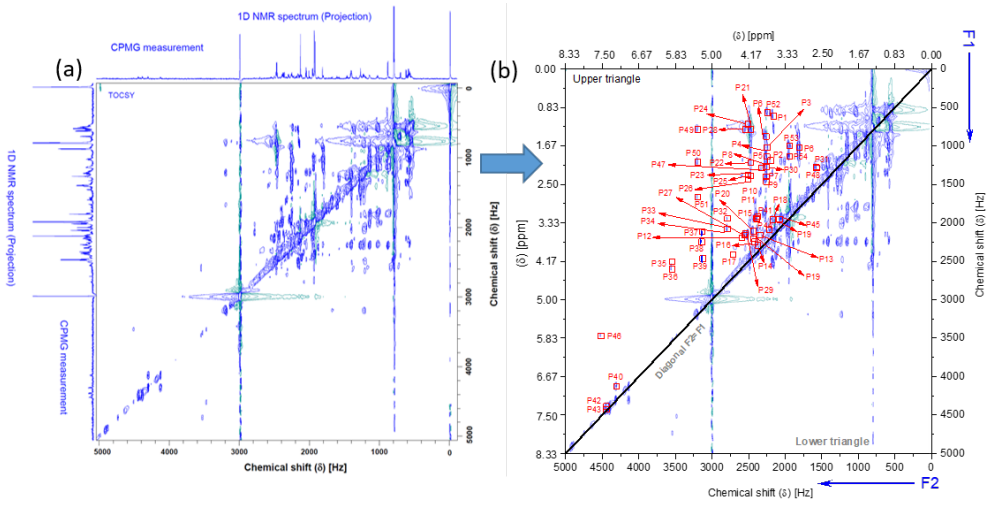


Figure 5.1 : (a) The ^{1}H-^{1}H TOCSY spectrum of a breast cancer tissue sampleat 600.13 MHz with τm of 80 ms. and relaxation time of 1 s, 16K and 128 data points acquired in the horizontal and the vertical dimension (F2, F1), resp. The NMR projections on F1 and F2 axes are an extra 1D NMR spectrum acquired by using the CPMG pulse sequence with excitation sculpting water suppression. (b) Peaks deduced from the experimental 2D TOCSY NMR spectrum.

5.3.1. Metabolites comprising the training dataset of breast cancer tissue, their frequencies and ppm in 2D NMR spectra.

Table 5.1 contains the chemical shift of horizontal and vertical frequencies and the corresponding metabolites in the breast cancer breast tissue. The dataset is available in https://doi.org/10.5281/zenodo.5724057.

Table 5.1: Breast cancer-tissue sample metabolites.

Metabolite	Metabolite [#]	F2 [Hz]	F1 [Hz]	F2 [ppm]	F1 [ppm]
3-Hydroxybutyrate	1	2496	1388.4	4.16	2.31
3-Hydroxybutyrate	1	2496	1448.4	4.16	2.41
3-Hydroxybutyrate	1	2496	722.4	4.16	1.20
Alanine	2	2256	876	3.76	1.46
Alfa-Glucose	3	3130.3	2112	5.22	3.52
Alfa-Glucose	3	3130.3	2224.7	5.22	3.71
Alfa-Glucose	3	3132	2568.5	5.22	4.28
Ascorbate	4	2405.3	2241.5	4.01	3.74
Ascorbate	4	2240.9	2064.4	3.73	3.44
Aspartate	5	2332.1	1590.9	3.89	2.65
Aspartate	5	2332.1	1681.6	3.89	2.80
Beta-Glucose	6	2778.6	1938.4	4.63	3.23
Beta-Glucose	6	2778.6	2084.3	4.63	3.47
Beta-Glucose	6	2778.6	2081.9	4.63	3.47
Glutamate	7	2248.2	1403.4	3.75	2.34
Glutamine	8	2258.4	1468.2	3.76	2.45
Glutamine	8	2258.4	1278	3.76	2.13
Glutathione	9	1529	1295	2.55	2.16
Glutathione	9	2262.5	1295	3.77	2.16
Inosine	10	3640.4	2567.4	6.07	4.28
Inosine	10	3640.4	2664	6.07	4.44
Isoleucine	11	2194.2	1181.4	3.66	1.97
Lactate	12	2462.9	790.4	4.10	1.32
Leucine	13	2231.4	1020.6	3.72	1.70
Lysine	14	1806	1032	3.01	1.72
Lysine	14	2250	1032	3.75	1.72
Lysine	14	2250	1137	3.75	1.89
Methionine	15	1578.3	1308.3	2.63	2.18
Methionine	15	2310.5	1308.3	3.85	2.18
Myo-Inositole	16	2112.5	1959.4	3.52	3.26
Myo-Inositole	16	2167.1	1959.4	3.61	3.26
Myo-Inositole	16	2429.9	2112.5	4.05	3.52
O-Phosphocholine	17	2571.6	2186.9	4.29	3.64
O-phosphoethanolamine	18	2408.9	1944.4	4.01	3.24
Phenylalanine	19	2390.3	1970.1	3.98	3.28
Phenylalanine	19	4453	4394.3	7.42	7.32
Phenylalanine	19	4453	4422.5	7.42	7.37
Proline	20	2471.9	1213.2	4.12	2.02
Proline	20	2471.9	1402.2	4.12	2.34
Serine	21	2375.3	2300	3.96	3.83
sn-glycero-3-phosphocholine (GPC)	22	2342.3	2163.5	3.90	3.61

Metabolite	Metabolite [#]	F2 [Hz]	F1 [Hz]	F2 [ppm]	F1 [ppm]
sn-glycero-3-phosphocholine (GPC)	22	2587.8	2195.8	4.31	3.66
Taurine	23	2049.9	1949.7	3.42	3.25
Threonine	24	2545.2	2144.3	4.24	3.57
Threonine	24	2545.2	791	4.24	1.32
Tyrosine	25	4316.1	4139.7	7.19	6.90
Tyrosine	25	2362.1	1920.4	3.94	3.20
Uracil	26	4513	3474.8	7.52	5.79
Valine	27	2160.6	617.4	3.60	1.03

5.4. ADIPOSE TISSUE-DERIVED HUMAN MESENCHYMAL STEM CELLS

Adipose tissue-derived human Mesenchymal Stem cells (AT-derived hMSCs) were obtained from the Cell Therapy Center (CTC)/The University of Jordan. The sample belongs to consented healthy females in the age range of 35-43, donor's recruitment and sample collection were approved by the Institutional Review board University of Jordan (IRB: CTC/1-2020/04 and approved on 10.03.2020).
Details of sample preparation can be found in the Appendix.

5.4.1. High resolution 1D and 2D NMR experiments

The NMR measurements were performed at Leibniz Institute for Analytical Sciences – ISAS, Dortmund, Germany. For ^{1}H NMR profiling, 600 µL of deuterium oxide (D_2O) (sigma Aldrich) was added to the lyophilized metabolite, in addition to an appropriate concentration of 3-(trimethylsilyl) propionate-2,2,3,3-d4 (TSP) as an internal reference and mixed thoroughly. Later, samples were transferred into high resolution 5 mm borosilicate glass NMR tubes (Boro-600-4-8) (Deutero GmbH) NMR tube. The high resolution ^{1}H NMR spectra of the intracellular extracted samples in addition to two reference samples were acquired using broadband high resolution 600.13 MHz (B_0 = 14.1 T) NMR Bruker spectrometer (Avance III 600) and the room temperature NMR probe (BBO model-Bruker) at 279 K. Acquisition and processing of NMR spectra were achieved by using the software Bruker TopSpin 3.6. The 1D NMR spectra were acquired using the 90° single-pulse experiment (Bruker pulse sequence zg) with embedded excitation sculpting for water suppression. ^{1}H-^{1}H TOCSY was acquired employing the phase-sensitive TOCSY experiment, using z-axis decoupling in the presence of scalar interactions (DIPSI)-2 spin-lock implemented in the Bruker pulse sequence *dipsi2esgpph*. The spectral range was set to 7 kHz in both dimensions, 16K and 128 data points acquired in the horizontal and the vertical dimension (F2, F1), respectively. Before 2D Fourier Transform, zero filling was performed to 32K and 1K data points in the horizontal and the vertical dimension, respectively. The spectral widths in the two dimensions were 12.00 ppm.

5.4.2. Metabolic Profiling Assignment

Metabolic assignment was accomplished using BMRB [176], HMAB [172] and Chenomx NMR Analysis Software. As a result, 32 metabolites were identified and annotated in the 1D spectra as shown in Figure 5.2.The spectra were referenced to the 2D contour of TSP,

base levels were equalized to eliminate background noise. Later, automated peak picking at a proper threshold was performed by applying the automatic method using the pp2 function in TopSpin 3.6, and then the obtained F2 and F1 frequencies were deduced.
In agreement with the 1D spectra, a total of 32 metabolites were assigned from the 2D NMR spectra as listed in Table 5.2. It can be observed that some metabolites appear and disappear during the cultivation and differentiation of the cells. NP in Table 5.2 exposes the disappearance of metabolites during the dynamic evolution of the cells. Looking at the obtained metabolic 1D and 2D NMR spectra, metabolic changes occurred in MSCs in response to prolonged cultivation. Differentiation is noticeable and mainly found in their lipid profiles. Multiple peaks are usually related to fatty acids that are normally produced by adipocytes that are predominant in the 1D and 2D NMR spectra of prolonged cultivated cells. MSCs differentiation is related to remodeling of lipidomic metabolism because different functional phenotypes are correlated with changes of the cellular membrane. [197, 214-216]. Beside fatty acids, myo-inositol (MI), taurine (Tau) and 1-methylnicotinamide (1-MNA) were not observed early in MSCs, however they were observed later in all MSCs groups by both 1D and 2D NMR spectra. Due to the variation in concentration of intracellular metabolites, the contour intensities of all TOCSY spectra were equalized (normalized to specific minimum threshold intensity) which was led to the disappearance of shallow peaks (the signal to noise ratio (SNR) < 3) as shown in Figure 5.2.

5.4.3. Intracellular metabolites detected in AT-derived hMSCs.

The chemical shift and the horizontal and vertical frequencies of metabolites in AT-derived hMSCs cultivated and differentiated under different conditions described in Chapter 8 are listed in Table 5.2. The dataset is available in https://doi.org/10.5281/zenodo.7276518.

5.5. DATA REPRESENTATION

In our datasets, each metabolite is represented by two main characteristic features of the 2D TOCSY spectra: the chemical shift frequencies on the horizontal and vertical axes. Since sufficient data samples is a vital element for classification, data augmentation is implemented to overcome the small datasets due to limited NMR data [217, 218]. Data augmentation is implemented to extend the number of data samples by simulating anticipated deviation on the original samples [219]. Thus, data augmentation results in duplicates of the samples, and the classifiers will deal with the same sample in different versions [220]. Data augmentation has been applied in spectrum classification in NMR [221], Raman spectra [219], and infrared spectra [222].

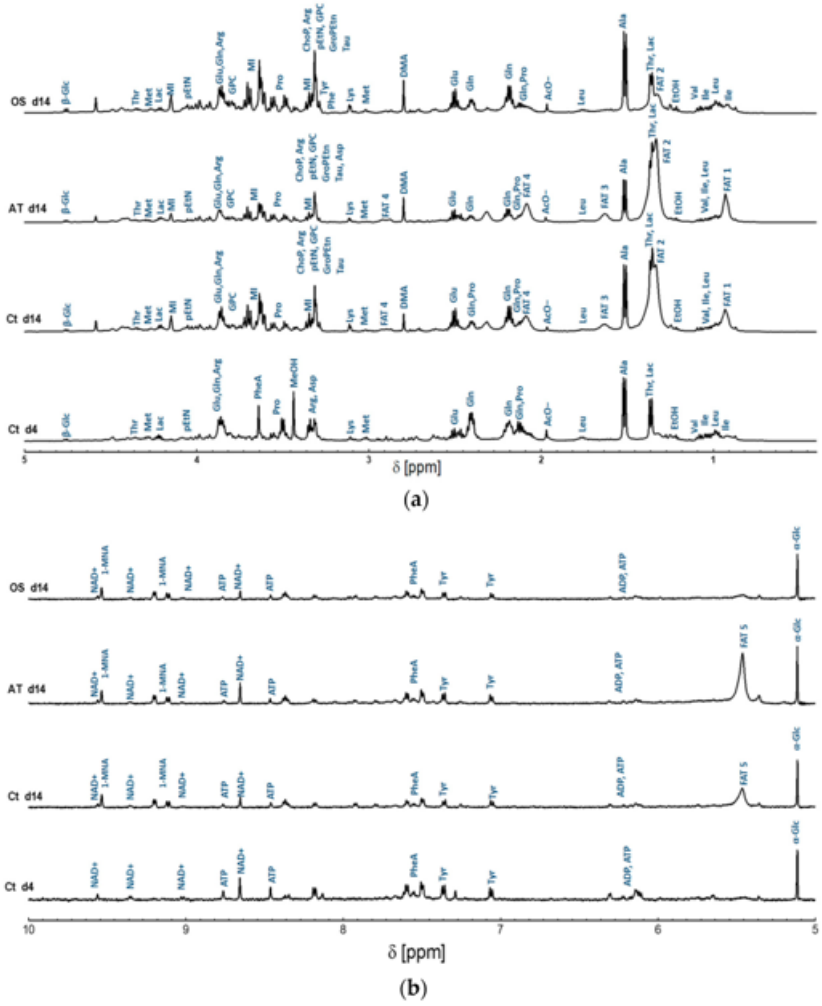


Figure 5.2: Representative high resolution 1H NMR spectra of intracellular metabolite extracts obtained from AT-derived hMSC samples collected at day 14 of differentiation into adipocytes and osteocytes, and their control samples represented in AT-derived hMSC collected at day 4 and 14 of cultivation in BCM. (a) 0.4–5 ppm region; (b) 5–10 ppm region. Peak assignment: Ile: Isoleucine; Leu: Leucine; Val: Valine; Thr: Threonine; Lac: Lactate; Ala: Alanine; Glu: Glutamine; Gln: Glutamate; Pro: Proline; Met: Methionine; Lys: Lysine; Arg: Arginine; GPC: Glycerophosphorylcholine; α-Glc: Alfa-Glucose; β-Glc: Beta-Glucose; MI: myo-inositol; ChoP: O-Phosphocholine; pEtN: Phosphorylethanolamine; GroPEtn: Glycerophosphorylethanolamine; ATP: Adenosine triphosphate; ADP: Adenosine diphosphate; Tyr: Tyrosine; Phe: Phenylalanine; NAD+: Nicotinamide adenine dinucleotide; Tau: Taurine; Asp: Asparagine; 1-MNA: 1-methylnicotinamide; AcO-: Acetate; DMA; Dimethylamine. In addition to the fatty acids signals; namely FAT 1, FAT 2, FAT 3, FAT 4, and FAT 5, representing methyl group -CH3, Acyl chains-(CH2) n-, methylene group -CH2-CH=CH, vinyl hydrogen -CH=CH, and diallyl methylene group =CH-CH2-CH=, respectively. The presence of ETOH (ethanol) and MeOH (methanol) was observed to represent residues from the cleaning and extraction procedures.

Table 5.2: Intracellular metabolites detected in AT-derived hMSCs at to control group at 4 days cultivation (Ct d4), 14 days of cultivation (Ct d14), 14 days of differentiation into adiobocytes (AT d14) and osteocytes (OS d14) and the standard frequencies from online libraries.

Metabolite	Ct d4		Ct d14		AT d14		OS d14		Standard	
	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1
Leu	752	494	762	484	762	489	772	494	720	540
Leu	950	590	930	584	930	589	934	560	900	600
Leu	900	720	914	720	913	750	910	760	900	720
Leu	1040	610	1073	631	1060	638	1046	608	1080	600
Leu	1040	695	1073	705	1060	709	1046	749	1080	720
Leu	1040	893	1073	923	1060	912	1046	922	1080	900
Leu	2229	611	2210	608	2212	610	2198	620	2220	600
Leu	2170	737	2171	750	2171	752	2163	749	2220	720
Leu	2163	943	2168	943	2157	908	2175	921	2220	900
Ile	1020	540	1020	540	1020	540	1020	540	1020	540
Ile	1023	408	1009	403	1047	417	1045	417	1020	600
Ile	2180	570	2170	580	2178	578	2220	540	2220	540
Ile	2160	1032	2165	1042	2195	1052	2210	1003	2220	1020
Tyr	1920	1778	1920	1787	1920	1782	1920	1790	1920	1830
Tyr	2396	1788	2253	1784	2355	1786	2358	1780	2340	1830
Tyr	2304	1877	2253	1848	2356	1836	2356	1848	2362	1920
Tyr	4073	4067	4090	3946	4094	3961	4095	3952	4316	4139
Phe	2340	1853	2354	1906	2261	1778	2360	1848	2390	1868
Phe	2340	1934	2254	1960	2254	1926	2260	1913	2390	1970
Phe	4362	4193	4362	4193	4368	4193	4370	4197	4453	4422
Phe	4362	4273	4362	4275	4368	4273	4370	4286	4453	4394
Glu	1373	1102	1354	1106	1354	1106	1349	1106	1470	1260
Glu	2337	1043	2337	1057	2344	1048	2333	1062	2258	1278
Glu	2341	1278	2344	1269	2341	1288	2333	1278	2258	1468
Gln	1295	1100	1281	1100	1284	1071	1288	1100	1260	1200
Gln	1378	1220	1384	1200	1389	1210	1370	1230	1380	1200
Gln	2190	1269	2186	1288	2191	1288	2194	1288	2220	1200
Gln	2225	1370	2227	1367	2210	1380	2208	1369	2220	1380
Lys	1740	809	NP	NP	1736	783	NP	NP	1800	840
Lys	1844	893	NP	NP	1836	898	NP	NP	1800	900
Lys	1836	1062	NP	NP	1836	1058	NP	NP	1806	1032
Lys	2295	962	NP	NP	2290	962	NP	NP	2220	900
Lys	2282	1057	NP	NP	2278	1044	NP	NP	2250	1032
Lys	2282	1118	NP	NP	2286	1119	NP	NP	2250	1137
FAT 1	NP	NP	616	405	600	420	NP	NP	600	420
FAT 2	NP	NP	789	545	789	531	789	545	785	535
FAT 3	NP	NP	1230	614	1245	620	NP	NP	1260	600
FAT 3	NP	NP	1240	1080	1260	1080	NP	NP	1260	1050
FAT 4	NP	NP	1715	772	1705	778	NP	NP	1792	766
FAT 5	NP	NP	3139	607	3150	607	NP	NP	3180	540
FAT 5	NP	NP	3138	1052	3150	1052	NP	NP	3180	1080
FAT 5	NP	NP	3140	1217	3150	1219	NP	NP	3180	1260
Lac	2499	715	2494	709	2494	715	2494	720	2463	790

Metabolite	Ct d4		Ct d14		AT d14		OS d14		Standard	
	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1
Thr	2160	789	2160	790	2160	720	2160	720	2160	780
Thr	2578	789	2537	790	2582	720	2573	720	2580	780
Pro	1879	1238	1869	1230	1873	1238	NP	NP	1980	1200
Pro	2408	1246	2408	1234	2408	1238	NP	NP	2472	1213
Pro	2408	1438	2435	1448	2405	1439	NP	NP	2472	1402
Ala	2270	723	2295	696	2295	705	2291	701	2256	876
Val	1350	632	1383	619	1394	619	1383	619	1380	617
Val	1875	1237	1890	1259	1880	1244	1870	1240	2160	617
Met	1518	1187	1523	1197	1518	1177	NP	NP	1560	1260
Met	2338	1270	2342	1274	2351	1277	NP	NP	2340	1260
Met	2338	1370	2342	1367	2351	1380	NP	NP	2340	1320
pEtN	2317	1852	2305	1848	2331	1865	2307	1849	2430	1950
GroPEtn	2300	1791	2291	1781	2293	1791	2292	1791	2300	1791
ChoP	2454	1950	2455	1953	2458	1954	2454	1947	2572	2187
GPC	2127	1943	2121	1943	2124	1939	2123	1939	2160	1980
GPC	2552	2333	2535	2338	2544	2359	2533	2348	2580	2340
Arg	1144	1000	1140	975	1143	984	1146	980	1120	920
Arg	1910	960	1920	960	1944	988	1928	988	1920	960
Arg	1974	1134	1978	1134	1986	1115	1969	1130	1920	1140
MI	NP	NP	2039	1880	2044	1869	2036	1869	2040	1800
MI	NP	NP	2088	1882	2093	1872	2087	1866	2112	1959
MI	NP	NP	2154	1970	2159	1977	2152	1972	2167	1959
MI	NP	NP	2460	2156	2452	2140	2452	2149	2423	2113
Asp	2390	1574	NP	NP	2398	1578	NP	NP	2400	1800
Asp	1870	1750	NP	NP	1876	1768	NP	NP	1800	1740
Tau	NP	NP	2064	1809	2065	1812	2063	1812	2040	1980
α-Glc	3135	2119	3125	2139	3137	2132	3140	2112	3130	2112
α-Glc	3135	2238	3125	2254	3137	2263	3140	2280	3130	2224
α-Glc	3135	2573	3125	2558	3132	2562	3140	2565	3130	2568
β-Glc	2760	1937	2774	1928	2765	1931	2759	1936	2778	1938
β-Glc	2717	2055	2714	2065	2717	2063	2717	2068	2778	2084
β-Glc	2717	2008	2714	2000	2712	2002	2714	2089	2778	2081
ATP	3620	2587	3640	2581	NP	NP	NP	NP	3620	2587
ATP	3620	2680	3640	2628	NP	NP	NP	NP	3620	2680
ADP	3569	2496	3566	2503	3566	2501	3570	2498	3569	2496
ADP	3569	2700	3566	2706	3569	2708	3569	2690	3569	2700
ADP	3569	2762	3566	2759	3569	2769	3569	2765	3569	2760
ADP	3569	2882	3566	2885	3569	2870	3569	2868	3569	2880
NAD+	5310	5110	NP	NP	5302	5106	NP	NP	5200	5110
1-MNA	NP	NP	5218	4718	5218	4725	NP	NP	5341	4921
1-MNA	NP	NP	5412	4718	5412	4725	NP	NP	5581	4921
1-MNA	NP	NP	5520	5328	5512	5321	NP	NP	5581	5341

An example of the data augmentation procedure for tyrosine is shown in Table 5.3. In SSL, before starting the classification process, data augmentation is used to create four disjoint datasets, training validation, learning, and testing sets. Each dataset will have

1200 data instances. In the training dataset, white Gaussian noise is added to the original frequencies with a different random signal-to-noise ratio. In the learning set, random noise is added to each instance of the original dataset. The validation and testing datasets are created by shifting the horizontal and the vertical frequency by a random value under a predetermined chemical shift constraint, within 30 Hz, 0.049 ppm, which is sufficient to simulate chemical shift fluctuations due to the NMR environmental matrix change [223].

Table 5.3: A subset of the training dataset showing the output of the data augmentation procedure for tyrosine.From one standard chemical shift for a metabolite, multiple versions of the same metabolite can be created.

Metabolite	**Standard From J-coupling**		**Experimental TOCSY**		**Augmented Generated**	
	F2 [Hz]	**F1 [Hz]**	**F2 [Hz]**	**F1 [Hz]**	**F2 [Hz]**	**F1 [Hz]**
Tyrosine	2353.3	1914.4	2362.1	1920.4	4317.1	4138.5
	4302.9	4138.3	4305.9	4139.3	4305.9	4139.0
					4315.3	4140.3
					2363.3	1921.7
					2361.4	1920.9
					2362.9	1919.1
					.....	.....

In ND scenarios, the training, validation, and testing datasets are used. Figure 5.3 shows the feature space of the metabolites in the breast cancer tissue sample. It can be observed that the frequencies overlap in the horizontal and vertical axes and cannot be linearly separated training dataset and adding random Gaussian noise to create the validation dataset [205, 219]. Data augmentation is applied on "control group at 4 days cultivation (Ct d4)" to create the training dataset. The training data set is of size $4000x2$, where 4000 is the number of independent samples from all existing metabolites on "Ct d4" and 2 is the dimension of the data, representing the horizontal and vertical frequencies. Due to the different number of multiples per metabolite, an uneven distribution of classes in the training dataset is observed and a class imbalance problem can arise. To overcome this issue, under-sampling of metabolites with more than two multiples has been applied during the data augmentation procedure. Figure 5.4 shows the feature space of the metabolites in Ct d4, Ct d14 (control group at 14 days of cultivation), AT d14 (after 14 days of adiobocytes differentiation) and OS d14 (after 14 days of osteocytes differentiation). It can be observed that peaks overlap on the horizontal and vertical axes and cannot be linearly separated. Similarly, in AT-derived hMSCs samples, multiple versions of the same metabolite are created by shifting the experimental chemical shift right and left up to 50 Hz to create the peaks overlap on the horizontal and vertical axes and cannot be linearly separated.

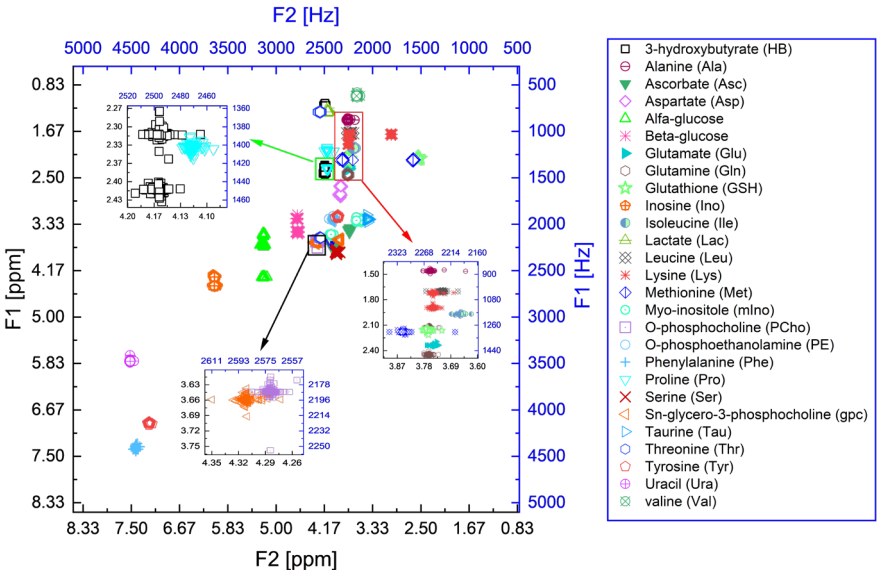


Figure 5.3: The feature space of the 27 metabolites deduced from the TOCSY spectrum of a breast cancer tissue. The magnifications are selected enlargements of peaks that overlap in (F1, F2) dimensions.

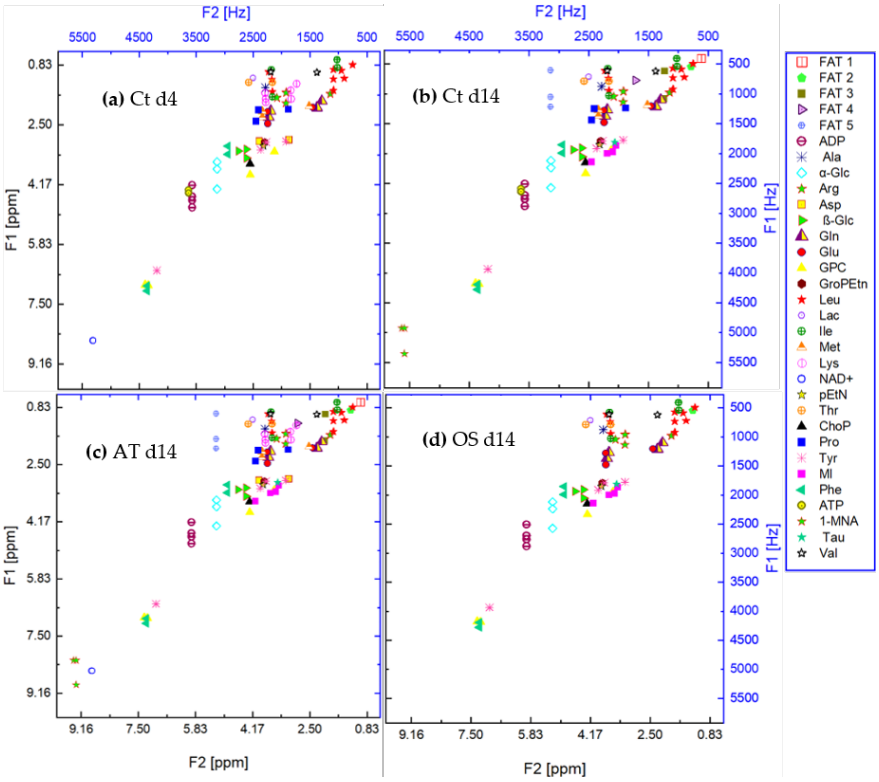


Figure 5.4: Feature space of the cross peaks of the metabolites contained in the samples (a) Ct d4, (b) Ct d14, (c) AT d14 and (d) OS d14. Abbreviations: Ile. Isoleucine, Leu. Leucine, Val. Valine, Thr. Thrionine, Lac. Lactate, Ala. Alanine, Glu. Glutamine, Gln. Glutamate, Pro. Proline, Met. Methionine, Lys. Lysine, Arg. Arginine, GPC. Glycerophosphorylcholine, α-Glu. Alfa-Glucose, β-Glc. Beta-Glucose, MI. myo-Inositol, ChoP. O-Phosphocholine, pEtN. Phosphorylethanolamine, GroPEtn. Glycerophosphorylethanolamine, ATP. Adenosine triphosphate, ADP. Adenosine diphosphate, Tyr. Tyrosine, Phe. Phenylalanine, NAD+. Nicotinamide adenine dinucleotide, Tau.Taurine, Asp.Aspargine and 1-MNA. 1-methylnicotinamide. In addition to FAT 1 to FAT 5.

6. Contribution: Semi-Supervised Learning in Metabolomics employing 2D TOCSY Spectrum

6.1. Semi-supervised Polynomial Classifier 45
6.2. Semi-supervised Support Vector Machines 45
6.3. Semi-supervised Kernel Null Foley–Sammon Transform 47
6.4. Experiments 51
6.5. Results and Discussion.....54
6.6. Validation.....56
6.7. Conclusion 60

Extreme peak shift and peak overlap are the main difficulties in metabolic profiling of NMR spectra of biofluid samples and tissues. Peak shifting aggravate the process of peak assignment of the same metabolites across various samples [13].

Using supervised learning, a classifier must be trained over the interval of possible chemical shifts for each metabolite, together with its multiplets all different pH values, concentrations, and temperature to reach an acceptable recognition rate. In cases where there is a shortage in the availability of training data, in addition to possible variations in data volume, supervised learning cannot be used efficiently. This situation is valid in NMR experiments due to the inapplicability to capture all settings in the dynamic environment of metabolites. In SSL, a training model is created based on a small, labeled amount of data which has been labeled by an expert or through supervised-learning scenarios. Later, the classifier is updated using the trained model together with the machine-labeled data [41]. SSL decreases the effort of capturing and adapting to all possible variations of different metabolites and can be a replacement for the manual assignment of metabolites in 2D NMR spectra. In this chapter, Polynomial Classifier (PC), Support Vector Machines (SVM) and Kernel Null Foley–Sammon Transform (KNFST) are introduced under the semi-supervised learning scenario. These classifiers are non-linear, which means that they map the original features of the dataset into a higher space, which might help in producing acceptable separability.

As discussed in Section 2.4.5, self-learning [224] is a subclass of the SSL methodology and can be used as a wrapper for different types of classification algorithms [41]. In self-training methods, the classifier itself is used to iteratively label or reject samples which belong to a larger unlabeled dataset. If not rejected, a sample together with its label, is added to the labeled dataset. Adding mislabeled data to the training dataset will have an undesirable effect on the classifier performance, therefore, adding only informative and certain predictions to the training set is an essential factor. These informative and certain predictions can be employed by introducing confidence bands, which are used to reject

possible outliers, i.e., do not lie in the confidence band threshold. Samples that do not exceed this threshold will be added to the training set and the classifier is retrained using the accepted data. The integration of SSL and confidence bands was used in field of traffic signs, handwritten digits [69, 73, 76], lunar elemental abundances [225] and gesture recognition [68]. Other confidence measures such as the mutual agreement and majority voting between multiple classifiers [226], uncertainty sampling [227] or conditional random fields [228, 229] were discussed in [45, 230]. The stopping conditions of the self-training technique are defined through one of the following measures: the maximum number of iteration is reached by the classifiers, the whole unlabeled set, is added to the labeled set or when there are no more confident predictions which can be added to the labeled set [45]. This chapter has been adapted and/or adopted from [205].

6.1. SEMI-SUPERVISED POLYNOMIAL CLASSIFIER

The Polynomial Classifier (PC) is a parameterized non-linear interpolation which transforms a sequence of input vectors to a higher dimension. PC has the form of an algebraic polynomial of order n. Let $N = \{1 \dots k\}$ be the number of training samples X, where $X = \{x_1, \dots, x_k\}$ of C different classes and class labels $y = \{y_1, \dots, y_k\}$. The polynomial discriminant function takes the form [72]

$$g(x) = {A_{PC}}^T \varphi(x) \tag{6.1}$$

where $\varphi(x)$ is the polynomial structure that represents all the possible multiplicative combination of the original feature X depending on the order of the polynomial n and on the dimension of the input vector [72]. The coefficient/weight matrix ${A_{PC}}^T$ is obtained during the training phase and is employed during the learning process to obtain the probability that a given feature belongs to class c. The polynomial discriminant function $g(x)$ creates a mapping from the feature space to a decision dimensional space that produces an output of posterior probability estimate to determine the class membership [72]. The solution of the model can be found using least squares optimization through minimizing the residual $\|{A_{PC}}^T \varphi(x) - g^*(x)\|^2$, where $g^*(x)$ is the optimal classification function [72].

Moore-Penrose pseudo-inverse approximation $\varphi(x)^+ = \left(\varphi(x)^T \varphi(x)\right)^{-1} \varphi(x)^T$ is used to estimate the model parameters ${A_{PC}}^T = \varphi(x)^+ g(x)$ during the training phase [34].
In the learning phase, the estimated weight matrix ${A_{PC}}^T$ is used to find the label of the new sample [34, 72, 76]. The number of free parameters N_{pc} in the confidence bands calculation is computed according to $N_{pc} = (L-1)M$, where L and M are the number of classes and the number of terms in the polynomial function [76]. In this work, we implemented third and fourth-order polynomial classifiers [72].

6.2. SEMI-SUPERVISED SUPPORT VECTOR MACHINES

The goal of Support Vector Machines (SVM) is to find a function with a maximum deviation from a target value f_{SVM} from the training data [231]. The original features are mapped into a higher dimensional space using a mapping function to find a hyperplane that separates the features. The support vectors are training samples which act as decision

boundaries to determine an optimal hyperplane that has the maximal distance to the nearest support vectors [34]. Let N be the number of training samples, X = $\{x_1, \dots, x_k\}$ are the features of the training samples with the labels $Y = \{y_1, \dots, y_k\}$, $\in \{-1, +1\}$. SVM finds a hyperplane that separates these classes by solving.

$$f_{SVM}(x) \rightarrow \omega_{SVM}^{T} \varphi(x) + b \qquad (6.2)$$

where φ is high-dimensional non-linear mapping of the features X, ω is the coefficient matrix and b is the bias vector. The hyperplane is optimized during the training phase by finding ω and b which maximize the distance between the support vectors and the hyperplanes [34]. In the learning phase, only Eq. (6.2) must be computed for every new instance. The implicit features mapping $\varphi(\vec{x}): \Re^n \rightarrow F$, where F is a high dimensional inner-product space, can be used to define kernel function $K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$ [34]. There is a wide range of kernels that can be used, Bishop [34] presents different kernels and discusses different conditions for constructing kernel functions. Throughout this work, the Gaussian kernel, or the radial basis function (RBF): $K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$ is used, where γ controls the bandwidth of the kernel function[232].

SVM is a binary classifier, i.e., a classifier tries to distinguish between two classes and the class membership is assigned according to the sign of label y. To solve the multi-class problem, SVM has to be reformulated to multiple binary problems and solved by combining these multiple binary classifiers. One approach is to use 'one-*vs*-all' classification. In this method, a multi-class problem is treated as multiple binary-classifiers in which a model is created using one class against all other classes. Suppose we have n classes $C = \{c_1, c_2, \dots c_{n-1}, c_n\}$ for class c_1, we consider c_1 as one class and all other classes $c_2 .. c_n$ are considered as another class. We build SVM model for class c_1. This procedure is repeated n times resulting in n models for n classes. The n multiple binary classifiers are then combined to create a multi-class classification problem. The label assignment for a new sample employs all n SVM models and assigning the label for the model with the highest output value [34, 233]. Another strategy is 'one-*vs*-one' in which $c(c-1)/2$ training models are built. An instance is classified according to a voting system [35]. On this work, the binary classification is extended into a multi-class approach by using one-*vs*-all classification.

Originally, SVM was designed as a classification problem where the label y is a discrete rather than a probability value. For comparing the degree of certainty of the prediction, obtaining a posterior class probability is useful. Several methods have been introduced to modify SVM to calibrate distance values into probabilities [53, 232, 234-236]. Platt [236, 237] fits the output of the SVM classification $P(y = 1|f_{SVM})$ using a sigmoid function with parameters A and B :

$$P(y = 1|f_{SVM}) \approx P_{A,B}(f_{SVM}) = \frac{1}{1 + exp\,(Af_{SVM} + B)} \qquad (6.3)$$

Platt defines a training set (f_{SVM_i}, t_i) where f_{SVM_i} and t_i are the output of the SVM classification and the target probability for training sample i respectively. The parameters $z^* = (A^*, B^*)$ are the optimal parameters to solve the maximum likelihood problem. The

number of positive samples N_+ and the number of negative samples N_- are used to describe the targets probability:

$$t_i = \begin{cases} \frac{N_+ +1}{N_+ +2} & \text{if } y_{i=} + 1 \\ \frac{N_- +1}{N_- +2} & \text{if } y_{i=} - 1 \end{cases} \quad , i = 1 \ldots . . l \tag{6.4}$$

In the sigmoid fit, instead of [0,1], the target probability t_i will be used and the sigmoid parameters are learned and estimated through minimizing the negative log likelihood of the training set with cross-entropy error [236, 237]:

$$\min_{A,B} \mathrm{f(z)} = -\sum_i t_i \, log(p_i) + (1 - t_i) \, log(1 - p_i) \tag{6.5}$$

$$\text{subject to } p_i = P_{A,B}(f_{SVM_i})$$

In this work, the tool box LIBSVM [232] is used. LIBSVM implements the extension of Platt and multi-class classification. Moreover, the confidence bands are calculated using Eq. (2.2), the degree of freedom v is defined as the difference between the total number of training samples and the number of support vectors [76] .

6.3. SEMI-SUPERVISED KERNEL NULL FOLEY–SAMMON TRANSFORM

Following [200, 201], let $\mathrm{X_c}$ denote the c^{th} class sample and $\mathrm{N_c}$ is the number of samples that belong to class c, then X is an n-dimensional sample with elements N belonging to c classes.

The within-scatter matrix $\mathrm{S_w}$, the between-class scatter matrix $\mathrm{S_b}$ and the total scatter matrix $\mathrm{S_t}$ are defined [238] as

$$\mathrm{S_b} = \sum_{i=1}^{C} \mathrm{N_i}(\mu_\mathrm{i} - \mu)(\mu_\mathrm{i} - \mu)^\mathrm{T} \tag{6.6}$$

$$\mathrm{S_w} = \sum_{\mathrm{i=1}}^{\mathrm{C}} \sum_{\mathrm{j=1}}^{\mathrm{N_i}} (\mathrm{x_i^j} - \mu_\mathrm{i})(\mathrm{x_i^j} - \mu_\mathrm{i})^\mathrm{T}$$

$$\mathrm{S_t} = \sum_{\mathrm{i=1}}^{\mathrm{C}} \sum_{\mathrm{j=1}}^{\mathrm{N_i}} (\mathrm{x_i^j} - \mu)(\mathrm{x_i^j} - \mu)^\mathrm{T}$$

Where $\mathrm{x_i^j}$is the jth sample that belongs to class i, $\mu = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{\mathrm{N_i}} \mathrm{x_i^j}$ is the sample mean, $\mu_\mathrm{i} = \frac{1}{\mathrm{N_i}}\sum_{j=1}^{\mathrm{N_i}} \mathrm{x_i^j}$ is the mean of the samples that belong to class i.

The Fisher Linear Discriminate criterion (FLD) finds the projection direction ω that best separates the data in terms of classification. FLD helps to find the linear transformation that yields a minimal within-class scatter and a maximal between-class scatter. Consequently, a sample is projected as close as possible to samples that belong to the same class and as far as possible to samples which belong to a different class. In terms of $\mathrm{S_w}$ and $\mathrm{S_b}$, FLD is defined [239] as:

$$\mathrm{J}(\omega) = \frac{\omega^\mathrm{T} \mathrm{S_b}\, \omega}{\omega^\mathrm{T} \mathrm{S_w}\, \omega} \tag{6.7}$$

Optimizing expression Eq.(6.7) using the generalized eigenvalue problem [238, 240], we get

$$S_b\,\omega = \lambda S_w \omega \tag{6.8}$$

By the definition of the generalized eigenvalue problem, S_w is non-singular and ω and λ are the generalized eigenvectors and corresponding eigenvalues of S_b and S_w. The eigenvalues are ordered such that $\lambda_1 \geq \cdots \geq \lambda_k \geq 0$ and the eigenvectors are orthonormal such that ${\omega_i}^t \omega_j = 0$ where $i \neq j$. Normalizing the eigenvectors such that $\|\omega_j\|^2 = {\omega_j}^T \omega_j = 1$ and collecting them in a matrix $\varphi = \omega^{(1)}, \ldots, \omega^{(k)}$, we can calculate the discriminate vectors of FST by $y = \varphi^T x$ [200].

The Null-Foley–Sammon Transform (NFST) suggests that we can find some null projection direction enforcing the conditions $\omega^T S_w\,\omega = 0$ and $\omega^T S_b\,\omega > 0$ in Eq.(6.7), so we get $J(\omega) = \infty$, such ω is called the Null Projection Direction (NPD) [201]. The best separability is ensured because all samples that belong to a given class are projected into one single point such that the within-class scatter is zero and at the same time, different classes are projected far from the rest of classes [201]. The idea of NFST is illustrated in Figure 6.1.

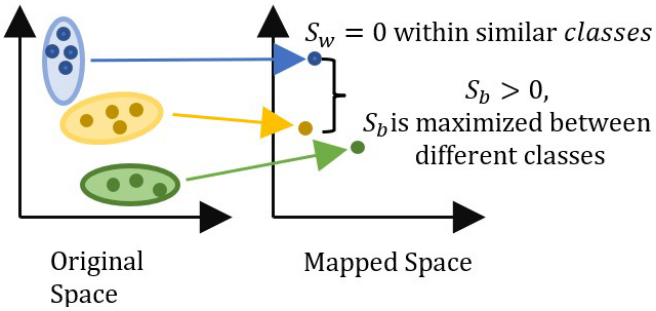


Figure 6.1: Geometrical visualization of NFST. Every class is represented by a single point in the mapped space. Test samples are mapped nearer to the class representation they belong to and far away from different classes.

The optimization problem for NFST [240] turns into :

$$J(\omega) = \max_{\omega} |\,\omega^T S_b\,\omega| \text{ subject to } |\,\omega^T S_w\,\omega| = 0 \tag{6.9}$$

To solve Eq.(6.9), we find $\omega \xrightarrow{\text{yields}} (\omega^T S_w\,\omega = 0 \wedge \omega^T S_b\,\omega > 0)$. It has been shown in [241, 242], that we can find the orthonormal basis B using Gram-Schmidt orthogonalization, and then we can write

$$\omega = \beta_1 b^1 + \cdots + \beta_m b^m = B\beta \quad \text{for each } \omega \in {Z_t}^{\perp} \tag{6.10}$$

Where ${Z_t}^{\perp}$ is the orthogonal complement of the null space of S_t. Then the solution β is computed through replacing ω by $B\beta$ in $\omega^T S_w\,\omega = 0$, and we can write [241, 242] :

$$(B^T S_w\,B)\beta = 0 \tag{6.11}$$

Solution β from Eq.(6.11) is used to compute the null projection direction ω using Eq. (6.10) to calculate the discriminant function NFST [241].

Let X_w be the matrix consisting of the vectors $x_i^j - \mu_i$ and X_t be the matrix consisting of the vectors $x_i - \mu$, we can define $S_w = \frac{1}{N} X_w X_w^T$ and $S_t = \frac{1}{N} X_t X_t^T$ [241], so Eq. (6.11) can be expressed as

$$HH^T \beta = 0 \quad \text{where } H = B^T X_w \tag{6.12}$$

The above Equation suggests that the eigenvalue problem solving the Null Space Discriminative direction is summed up to an inner product problem which proposes to extend the algorithm using kernels. Although NFST turns out to be a successful classifier [240], but due to its linear approach, it is inadequate to classify real- world example. Therefore, it is extended to perform classification in non-linear models using kernels. By incorporating kernels, Eq. (6.6) are rewritten as

$$S_b^{\varphi} = \sum_{i=1}^{C} N_i (\mu_i^{\varphi} - \mu^{\varphi})(\mu_i^{\varphi} - \mu^{\varphi})^T \tag{6.13}$$

$$S_w^{\varphi} = \sum_{i=1}^{C} \sum_{j=1}^{N_i} (\varphi(x_i^j) - \mu_i^{\varphi})(\varphi(x_i^j) - \mu_i^{\varphi})^T$$

$$S_t^{\varphi} = \sum_{i=1}^{C} \sum_{j=1}^{N_i} (\varphi(x_i^j) - \mu^{\varphi})(\varphi(x_i^j) - \mu^{\varphi})^T$$

Where μ^{φ} and μ_i^{φ} are the mean of all samples in the higher space and the mean of class *i* respectively. The fisher criteria in the higher space can be defined [240] as

$$J^{\varphi}(\omega) = \frac{\omega^T S_b^{\varphi} \omega}{\omega^T S_w^{\varphi} \omega} \tag{6.14}$$

The optimization problem can be written [240] as

$$J^{\varphi}(\omega) = \max_{\omega} \left| \omega^T S_b^{\varphi} \omega \right| \text{ subject to } \left| \omega^T S_w^{\varphi} \omega \right| = 0 \tag{6.15}$$

The orthonormal set in the mapped space can be found using kernel PCA [241, 243].The kernel PCA algorithm [241] uses the centralized kernel $\overline{K} = (I - 1_N)K(I - 1_N)$, where K is the kernel matrix of the mapped training data, I is the $N \times N$ identity matrix and 1_N is a $N \times N$ matrix with all elements equal to $\frac{1}{N}$. Applying the eigenvalue decomposition of $\overline{K} = VEV^T = \sum_{i=1}^{n} \lambda_i v_i v_i^T$, where V is the $N \times N$ matrix whose columns contain the eigenvectors v_i of $\overline{K}$ and E is a diagonal matrix containing the corresponding eigenvalues λ, where $\lambda_1 > \cdots > \lambda_n$. With $\overline{K}$ being guaranteed to be positive-definite [34] and V be an orthonormal matrix, we can define a factor matrix of the form $\breve{V} = VE^{1/2}$, which defines a scaled eigenvector that contains the coefficient for the normalized orthonormal basis that is to be replaced [241, 244] in Eq. (6.11). The orthonormal basis B_{new} can be expressed by the centralized data in kernel space $\overline{\varphi(x)}$ and coefficient vector

[241] $\breve{V}$ as $B_{new} = \overline{\varphi(x)}\breve{V}$. Let the matrix $H_K = \left((I - 1_N)\breve{V}\right)^T K(1 - L)$, where L a block diagonal matrix with block sizes equal to the class-specific number of samples N_c and the value $1/N_c$ of each non-zero element. The factor $(I - 1_N)$ is a normalization of the basis vector coefficient due to zero-mean mapping resulting from kernel centralization [241, 242]. Replacing H by H_K in (6.11), we obtain $\beta^1 \dots \beta^{c-1}$ solutions and we can calculate $c - 1$ projection directions $\overline{\omega}^j$ using the coefficient vector $\breve{V}$, $\overline{\omega}^j = \left((I - 1_N)\breve{V}\right)\beta^j \quad \forall j = 1, \dots, c - 1$. To find the null space projection for point z on $\overline{\omega}$, we calculate $\overline{\omega}^T K(z) = {\beta^j}^T \left((I - 1_N)\breve{V}\right)^T K(z)$ [241, 244].

The test point z is mapped to $(K(z)^T\overline{\omega}^1, \dots K(z)^T\overline{\omega}^{c-1})^T$, with $K(z)$ as the kernel function of sample z [240-242].

KNFST was used an outlier detection in previous work [241, 245, 246], nevertheless, in this work we have extended the functionality of KNFST to be employed in the SSL scenario as following: During the training phase, the projection direction ω, the class-wise projections of training data into the null space D [241], in addition to the confidence band for each sample is computed using the training data. During the learning process, for each sample $z_{unlabeld} \in X_{unlabeled}$, the projection z^* using ω is computed. The class membership is computed according to

$$Class(z^*) = \min_{1 \le i \le C} dist(z^*, D) \qquad (6.16)$$

In Eq. (6.16), the class membership $Class(z^*)$ is computed by calculating the Euclidean distance between the projected sample z^* and the projection of all classes in the mapped null space. The instance z^* is assigned to the nearest class as depicted in Figure 6.2. Next, the confidence band for z^* is computed according to Eq. (2.2).The degree of freedom for the t-student distribution is the difference between the size of the feature space and the size of projected dimension [247].

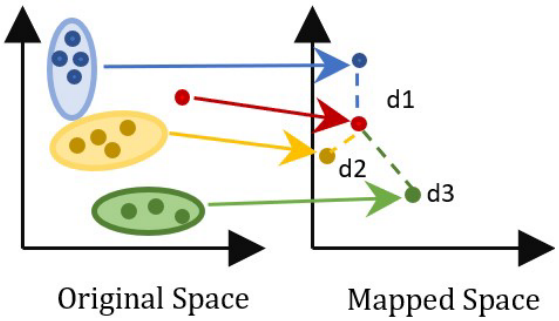


Figure 6.2: Class membership in KNFSTis determined according to the distance between the projected class and the new red sample. The blue, yellow and green classes are mapped into one point for each class in the mapped class. The assignment of the new red sample is determined according to the distance between its projection and the projection of the other classes (d1, d2, d3). Distance d2 is the shortest distance to the red class, therefore, it is more probable that the red sample belongs to the yellow class.

Initially, confidence bands are computed from the training data and their values are the main criterion to decide whether a sample is used to update the training set. A relative

deviation of the confidence value of training data is allowed, i.e., an unlabeled sample can be added to the training set once it is corresponding confidence value falls within this deviation. Once the sample is accepted, it is added to the training set together with its label as well as its confidence value. At last, the classifier is retrained after a maximum of t samples has been added to the training dataset. For the sample z^*, we construct a two-sided normalized confidence band $(\sigma_{min}, \sigma_{max})$ such that $probability((\sigma_{\min}, \sigma_{\max}) \ni \sigma_z) = 1 - \alpha$, where σ_z is the computed confidence band for sample z. The values of $\sigma_{\min}$ and $\sigma_{\max}$ are calculated as $\sigma_{\min} = quantile(\sigma_{\text{Train}}, \ell^{\min})$ and $\sigma_{\max} = quantile(\sigma_{\text{Train}}, \ell^{\max})$, where $\ell^{\max}$ and $\ell^{\min}$ are experiment-dependent and σ_{train} is the confidence band vector of the training data. Generally, all possible combinations values $0 < \ell^{\max} \leq 1$ and $0 < \ell^{\min} \leq 1$ could be examined [248]. In our settings, if multiple combinations of ℓ^{max} and ℓ^{min} achieve a similar accuracy and misclassification rate, then we choose the configuration with the narrowest confidence band. Figure 6.3 and Figure 6.4 summarize the steps in the training as well as in the learning phases of KNFST, respectively.

6.4. EXPERIMENTS

In the scenario of semi-supervised learning, a third (PC3) and fourth order (PC4) polynomial classifier, KNFST, and SVM classifiers are tested. The performance of the classifiers related to an increased size of the initial training set was investigated and plotted in Figure 6.5 to Figure 6.6. The learning procedure is repeated for different initial amounts of training data to examine the role of the size of the initial dataset on the learning process and to observe the minimum ratio of the initial training set, which is sufficient to produce an acceptable performance. The labeled dataset is partitioned into ten portions of training data. The system uses random initial training samples, starting from 10%, 20%, 30%, until reaching 100% of the training data. This random division and permutation of the training dataset will lead to a different number of samples per metabolite; this is important to monitor how classifiers will handle unbalanced datasets in diverse experimental situations. Therefore, it is essential to repeat the experiment multiple times and enforce the classifiers to deal with random permutation and partition to obtain accuracy expectations independent of the partition of the training dataset. The labeled dataset is partitioned into ten portions of training data. The system starts by using random initial training samples, starting from 10%, 20%, 30%, until reaching 100% of the training data size. For each portion of the initial training dataset, ten runs are performed. Thus, the classifiers will perform the experiments ten times for each of the ten partitions of the training dataset. A testing dataset, of size 1200×2, which is created using data augmentation is used to test the performance of the SSL scenario.

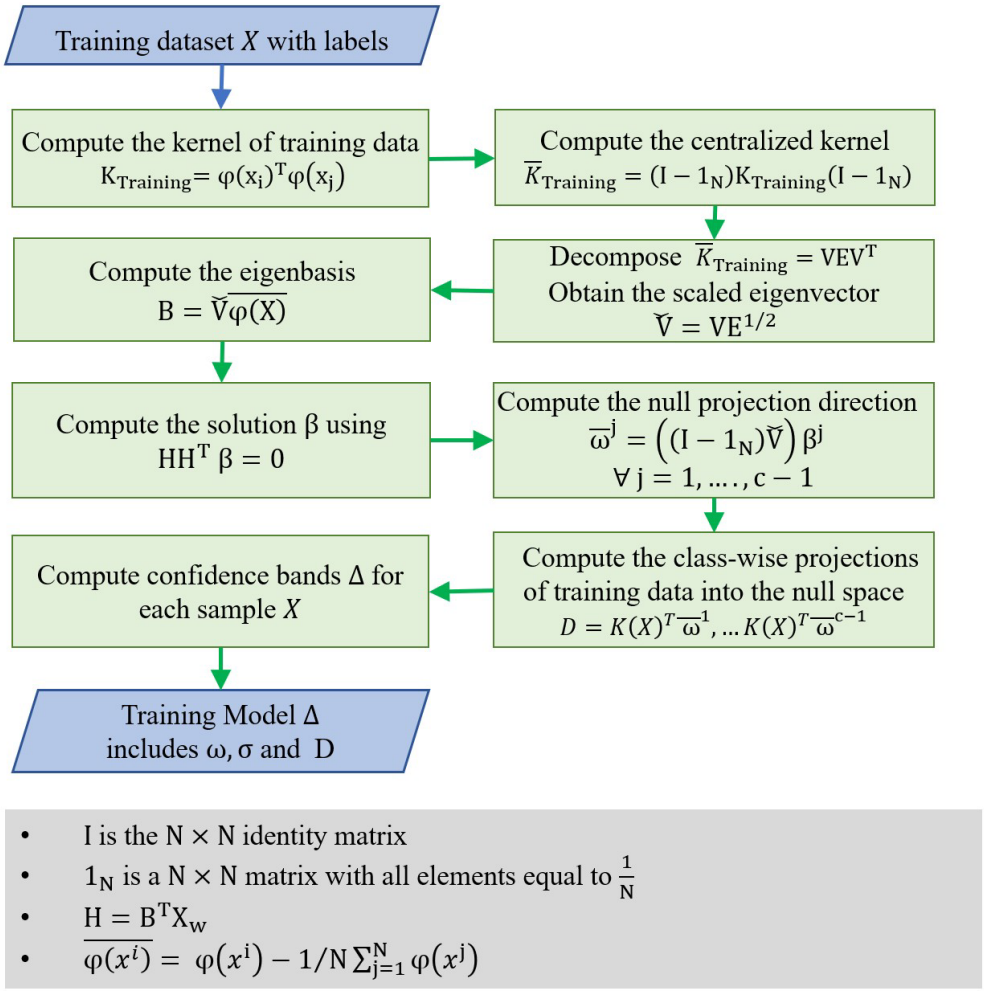
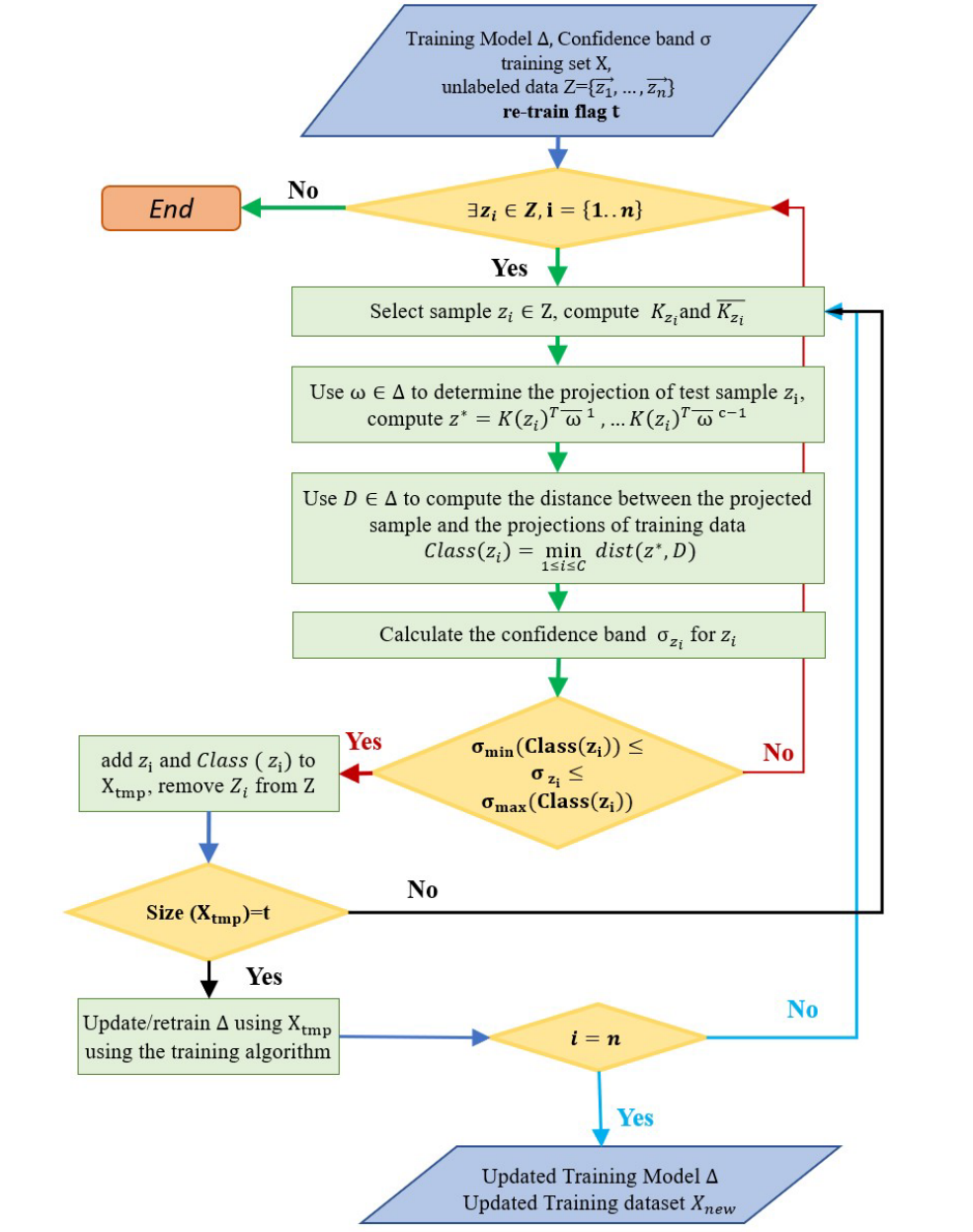


Figure 6.3: The training phase in semi-supervised KNFST algorithm. The aim of the training phase is to generate a training model based on training dataset. The training models consists of the optimized projection matrix, confidence bands values and the class-wise projections of training data into the null space.



- X_{tmp} are the accepted samples that will be added to the training dataset. X_{tmp} contains the confident predicted samples and their labels.
- Re-train flag t is the number of instance collected in X_{tmp} before retraining the classifiers.
- Class is the class label assigned to a sample.

Figure 6.4: The learning phase in Semi-supervised KNFST algorithm. The learning process starts by using the pre-generated training model. SSL iteratively selects a sample from the unlabeled data. The classifier predicts a label for the sample where new labels are accepted if the confidence band value is within a range $\sigma_{min} \le \sigma \le \sigma_{max}$. Those accepted samples are added to the initial training set together with their predicted labels after t accepted samples, where t is a re-train flag used to check the number of accepted samples before retraining the classifier. The classifier is retrained on those t samples, creating a new training model that will be used to predict the labels for the rest of the unlabeled data and new confidence bands are calculated. This procedure is repeated until no unlabeled data matches the confidence band conditions, if there is no qualified example left, the algorithm terminates.

The assessment of the results is based on the accuracy of the classification:

- $\text{Accuracy} = \frac{\text{Number of correctly classified samples}}{\text{Total number of samples}}$
- $\text{Mislabeling rate} = \frac{\text{Number of wrongly classified samples added to the training set}}{\text{Total number of learned examples added to the training set}}$

6.5. RESULTS AND DISCUSSION

The accuracy and the mislabeling of the classifiers versus the size of initial training data are displayed as boxplots of median and standard deviation for ten different processing runs. Figure 6.5a shows the classification accuracy of KNFST, SVM, PC3, and PC4 classifiers. From the plot, the accuracy of KNFST and SVM increases with an increasing initial amount of labeled data until reaching around 100% at the size of 20% of the initial training dataset, where it is corresponding at this point to only eight samples per metabolite.

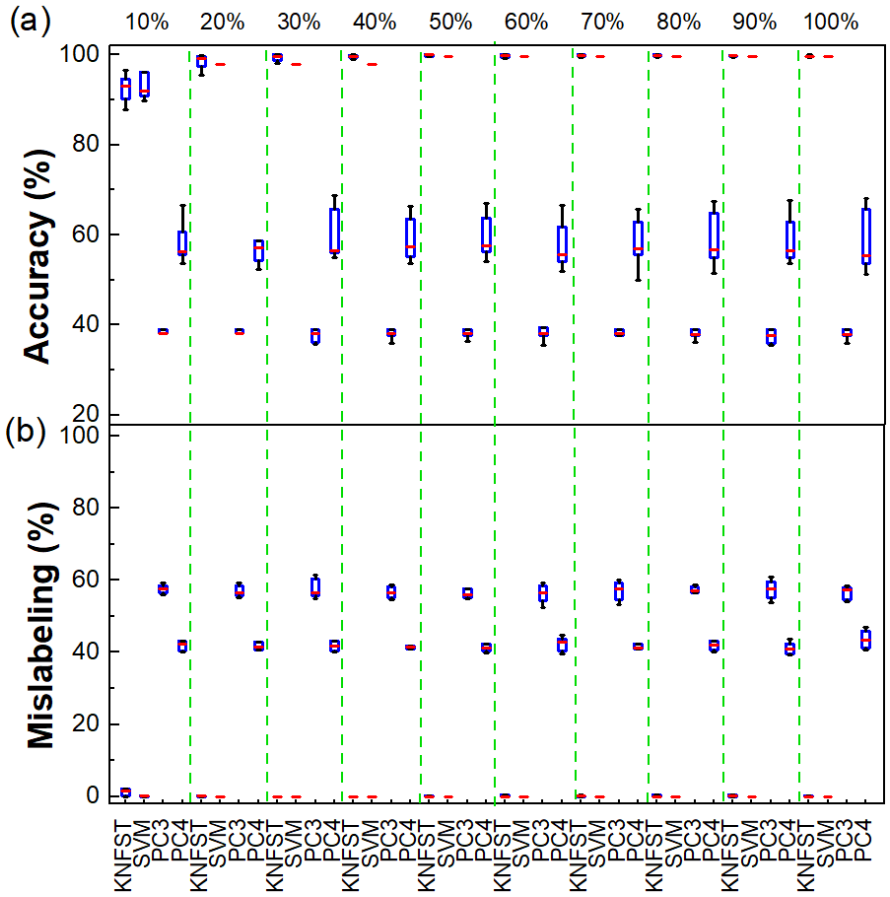


Figure 6.5: The accuracy and mislabeling versus different sizes of initial training data.

Conversely, PC3 and PC4 showed a lower accuracy in comparison and no improvement in the performance with the increasing size of the training dataset. The most probable explanation is the high mislabeling rate, shown in Figure 6.5b, where PC3 and PC4 have mislabeling rates of around 60% and 45%, respectively, overall sizes of the training dataset. Noticeably, both PC3 and PC4 were unable to learn any samples until using 30%

and 40% initial labeled training data. Remarkably, the mislabeling (misclassification) of KNFST and SVM starts with a rate of less than 5% (considered significantly low), and it decreases with increasing training set size reaching nearly 0%.

Analyzing the performance of the classifiers in the presence of an extremely small amount of initial training data, as low as one or two labeled samples per metabolite, is also noteworthy for this work since an NMR dataset is always kept as small as possible to save measuring time and to avoid sample alteration with time, leading to data scarcity.

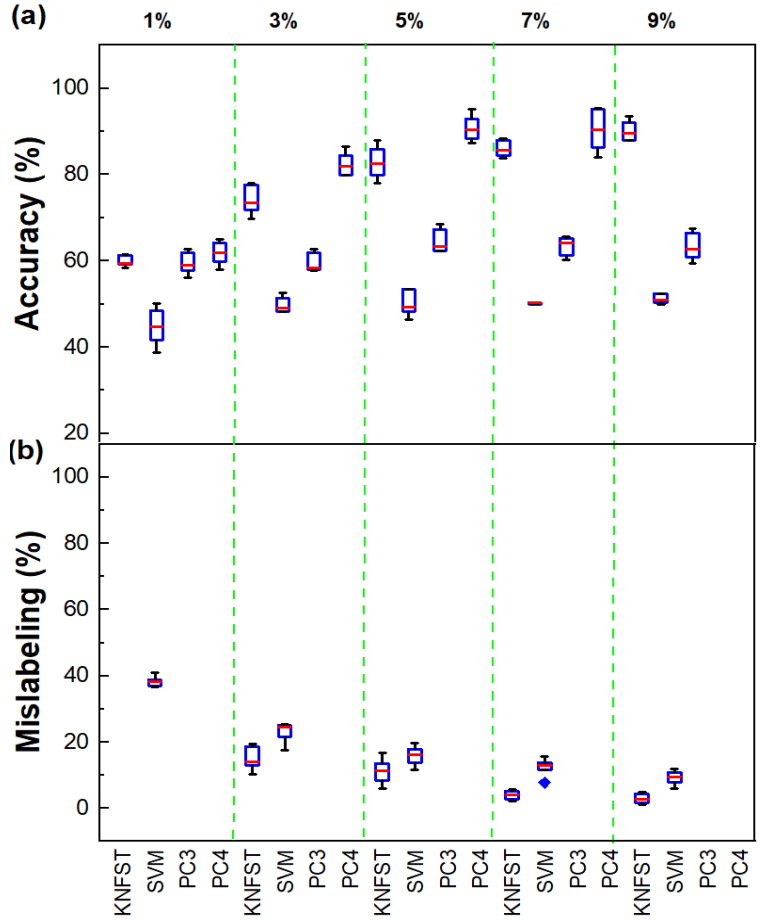


Figure 6.6: The accuracy and mislabeling versus different sizes of initial training data dataset for small initial amounts of labeled training data (≤9% of the entire dataset).

Figure 6.6a shows the accuracy of the classifiers in these cases with only 1% of the training dataset, ensuring one sample per metabolite per multiplet was the starting of the classification. Interestingly, the accuracy of SVM and KNFST kept increasing steadily despite the extremely small size of the initial training dataset. Additionally, the accuracies of both KNFST and SVM reached 90% at an initial training data set the size of 9%. The mislabeling rate of the SVM is around 40% at 1% of the initial training dataset of size, as shown in Figure 6.6b. No mislabeling rates appear for KNFST because it was not able to learn any sample. Later, the values of mislabeling of KNFST and SVM were around 15% and 25%, respectively. These values of mislabeling were decreasing with increasing initial training data set size. Within the low training data set size settings, KNFST showed a

higher performance than SVM, while both showed better accuracy than PC3 and PC4 at extremely low size settings. The mislabeling rates of PC3 and PC4 for extremely low sizes of the initial training data could not be defined (see Figure 6.6). This is typical for polynomial classifiers since they commonly require a relatively large amount of training data in order to be able to generalize [76]. It is essential that when a classifier is unable to learn any data samples and hence does not appear on the figures, the whole classification process turns into a supervised learning procedure rather than semi-supervised learning. This happens because no new data samples will be added to the initial training data set when the classifier does not learn any sample. Therefore, the test dataset will be tested against the un-updated original training data set. This explains the accuracies that appear in Figure 6.6a despite the absence of mislabeling in Figure 6.6b.

6.6. VALIDATION

The metabolite assignments of the breast cancer sample were validated based on the matching between the metabolites standard chemical shift from 1D NMR and 2D TOCSY with the experimental 2D TOCSY on the same sample (breast cancer tissue). Every metabolite 2D TOCSY standard chemical shift was deduced from the standard chemical shift 1D NMR from the Batman [13], BMRB [176], and HMDB [172] databases as well as relevant literature [249, 250].

Standard (F2, F1) cross-peak entries of ^{1}H-^{1}H TOCSY of the metabolites that appeared in the studied breast cancer tissue are listed in Table 6.1. Standard entries (indicated in the table) were deduced from the coupled peaks that appeared in standard 1D NMR spectra from affirmed databases as well as standard 2D TOCSY [3, 13, 172, 176, 249, 250]. Experimental cross-peaks are deduced from the measured TOCSY of the sample. Characteristic (F2, F1) cross-peak entries of every metabolite that has been used for the assignment are listed. These peaks are labeled with P1 until P48, and they are annotated in Figure 5.1b.

After the chemical shift verification of the cross-peak entries, the chemical shifts were assigned to metabolites. The results were verified and confirmed according to the published work on the same sample of the same scientific group [204, 251].

The demonstrated assignment in Figure 6.7 was done considering the results of the KNFST classifier only because it has shown the highest accuracy. The metabolite assignment was perfect (100%) without an occurrence of mismatching of the entries. Interestingly, the KNFST classifier matched all metabolites, although, for some metabolites, the chemical shift deviation was around 30 Hz (0.049 ppm), corresponding to a severe deviation that may cause substantial uncertainty in the metabolic assignment.

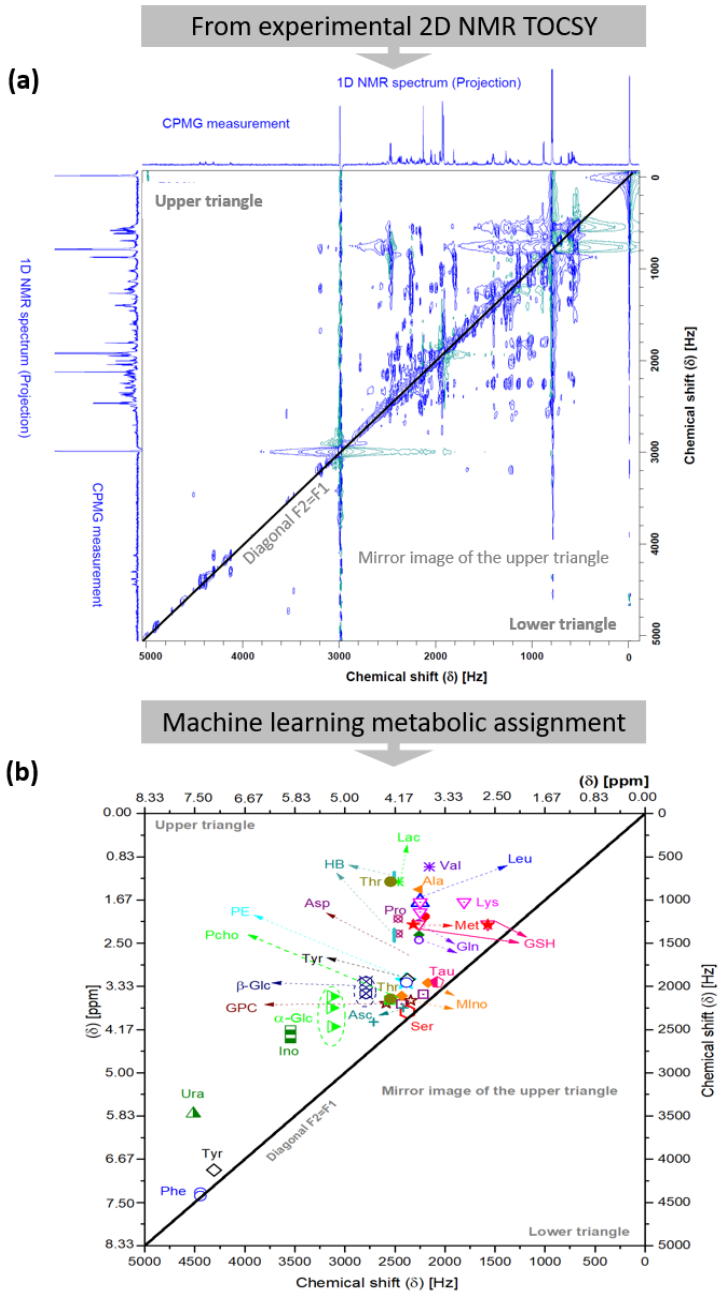


Figure 6.7: The metabolite assignment based on (a) the experimental 2D TOCSY NMR spectrum of breast cancer tissue after considering (b) the results of the KNFST classifier., which provides the highest accuracy. Acronyms of the metabolites are Val: Valine; Ile: Isoleucine; Leu: Leucine; Lys: Lysine; Glu: Glutamate; Ala: Alanine; Gln: Glutamine; Asp: Aspartate; GPC: sn-glycero-3-phosphocholine; Ser: serine; PE: O-phosphoethanolamine; Asc: ascorbate; mIno: myo-Inositole; Lac: Lactate; Pro: Proline; HB: 3-Hydroxybutyrate; PCho: O-Phosphocholine; Thr: Threonine; GSH: Glutathione; β-Glucose; α-Glucose; Ino: Inosine; Tyr: Tyrosine; Phe, phenylalanine; Tau: Taurine; Ura: Uracil; Met: methionine.

Table 6.1: Standard and experimental (F2, F1) Hz cross-peak entries of ^{1}H-^{1}H TOCSY of the metabolites appeared in the studied real breast cancer tissue.. Standard entries (indicated in the table) were deduced from the coupled peaks that appeared in standard 1D NMR spectra from affirmed databases [13, 172, 176, 249, 250]. Experimental (F2, F1) Hz cross-peaks are deduced from the experimental TOCSY measurement of the sample. Only characteristic (F2, F1) Hz cross-peak entries of every metabolite are listed, and they are labelled with P1 to P48 and annotated in Figure 5.1b.

		1D Spectra Peak Position	**Peak**	**Standard**		**Experimental**	
				From 1D NMR coupling		**From 2D TOCSY**	
#	**Metabolite**	[PPM]	*Posit-ion*	F2 [Hz]	F1 [HZ]	F2 [Hz]	F1 [HZ]
1	Valine	0.976, 1.029, 3.601	P1	2160.6	617.4	2159.4	615.4
2	Isoleucine	1.249, 1.458, 1.249, 1.969, 3.657, 0.927, 0.998	P2	2194.2	1181.4	2190.4	1182.2
3	Leucine	0.94, 0.95 3.719, 1.701	P3	2231.4	1020.6	2238.4	1020.2
4	Lysine	1.72, 3.01, 3.75 1.895	P4, P5	1806.0	1032.0	1812.3	1026.2
				2250.0	1032.0	2244.4	1026.2
				2250.0	1137.0	2244.4	1140.2
5	Glutamate	3.747, 2.078, 2.339	P6	2248.2	1403.4	2259.2	1404.3
6	Alanine	1.46, 3.76	P7	2256.0	876.0	2262.4	882.2
7	Glutamine	3.764, 2.13, 2.447	P8, P9	2258.4	1278.0	2262.4	1278.2
				2258.4	1468.2	2262.4	1464.3
8	Aspartate	3.886, 2.802, 2.651	P10, P11	2332.1	1590.9	2323.2	1602.2
				2332.1	1681.6	2323.4	1685.1
9	sn-glycero-3-phosphocholine (GPC)	3.605, 3.672,3.903, 3.871.3.946, 4.312,3.659, 3.212	P12, P13	2587.8	2195.8	2587.9	2210.5
				2342.3	2163.5	2367.8	2117.7
10	Serine	3.833, 3.958	P14	2375.3	2300.0	2390.2	2294.6
11	O-phosphoethanolamine	3.240, 4.014	P15	2408.9	1944.4	2390.4	1941.1
12	Ascorbate			2240.9	2064.4	2217.1	2090.0

		4.857, 4.771, 3.734, 3.440	P16, P17	2405.3	2241.5	2435.0	2204.1
13	Myo-Inositole	3.518, 4.049, 3.611, 3.265	P18, P19, P20	2112.5	1959.4	2076.8	1958.9
				2167.1	1959.4	2170.2	1958.9
				2429.9	2112.5	2432.1	2109.0
14	Lactate	4.104, 1.317	P21	2462.9	790.4	2468.2	787.5
15	Proline	4.119, 3.407, 3.323, 2.002, 2.080, 2.336, 2.022	P22, P23	2471.9	1213.2	2468.2	1217.7
				2471.9	1402.2	2468.2	1389.7
16	3-Hydroxybutyrate	4.160, 2.414, 2.314, 1.204	P24, P25, P26	2496.0	722.4	2506.6	718.4
				2496.0	1388.4	2506.6	1376.6
				2496.0	1448.4	2506.6	1438.7
17	O-Phosphocholine	4.285, 3.644	P27	2571.6	2186.9	2550.5	2161.1
18	Threonine	4.241, 1.318, 3.573	P28, P29	2545.2	791.0	2543.6	787.7
				2545.2	2144.3	2543.6	2143.4
19	Glutathione	4.557, 2.97, 2.943 3.766, 2.548, 2.158	P30, P31	1529.0	1295.0	1572.0	1277.7
				2262.5	1295.0	2260.7	1277.7
20	Beta-Glucose	4.630, 3.230, 3.473, 3.387, 3.450, 3.882, 3.707	P32, P33, P34	2778.6	1938.4	2788.3	1944.4
				2778.6	2084.3	2788.3	2083.8
				2778.6	2081.9	2788.3	2080.3
21	Inosine	8.189, 8.310, 6.066, 4.752, 4.439, 4.278, 3.882	P35, P36	3640.4	2567.4	3543.4	2501.8
				3640.4	2664.0	2868.5	2603.0
22	Alfa-Glucose	5.216, 4.630, 3.519, 3.698, 3.822, 3.826, 3.749	P37, P38, P39	3130.3	2112.0	3131.9	2115.7
				3130.3	2224.7	3140.2	2248.9
				3132.0	2568.5	3127.7	2464.9
23	Tyrosine	7.192, 6.898, 3.200,	P40, P41	23621	1920.4	2374.5	1920.4
				4316.1	4139.7	4307.3	4124.8

		3.055, 3.936					
24	Phenylalanine	3.283, 3.113, 3.983, 7.322, 7.420, 7.369	P42, P43 P44	4453.0	4394.3	4443.9	4387.0
				4453.0	4422.5	4443.9	4425.9
				2390.3	1970.1	2384.4	1954.8
25	Taurine	3.246, 3.410	P45	2049.9	1949.7	2078.7	1951.2
26	Uracil	5.79, 7.52	P46	4513.0	3474.8	4513.3	3471.7
27	Methionine	3.850, 2.183, 2.122, 2.629	P47, P48	2310.5	1308.3	2316.6	1285.1
				1578.3	1308.3	1571.4	1286.3

6.7. CONCLUSION

This work enabled the automatic and accurate spectral assignment of metabolites based on deconvolution of 2D-TOCSY NMR spectra by employing a semi-supervised machine learning approach. We have customized and extended four semi-supervised learning classifiers to test the automatic assignment under different initial training set sizes. The correctness of the metabolic assignments by our approach in applying 2D TOCSY spectra was based on comparing the results deduced from 1D-NMR spectra by human specialists on the same samples. The KNFST and SVM classifiers show high performance and low mislabeling rates for small and large sizes of the initially labeled training data set. To accept or reject the classification results of the classifiers, the concept of confidence bands was implemented. Under the same settings, both polynomial classifiers show a much weaker performance. For an extremely small size (≤9% of the entire dataset) of the initial training data set, PC3 and PC4 polynomial fail to provide satisfactory performance compared to KNFST and SVM classifiers, while the latter provided satisfactory performance as well as a low mislabeling rate. Hence, KNFST and SVM show superior performance over the other tested classifiers at every size of the initial training dataset. Our study demonstrates that machine learning in metabolite assignments based on the 2D TOCSY NMR spectra approach can be considered accurate and robust.

7. Contribution: Novelty Detection in Metabolomics Employing 2D TOCSY Spectrum

7.1. Kernel Null Foley-Sammon transform....63
7.2. Support Vector Data Description....63
7.3. Kernel Density Estimation....65
7.4. Threshold setting and novelty detection....65
7.5. Novelty detection of metabolites using breast cancer tissue....66
7.6. One-class novelty detection....67
7.7. Multi-class novelty detection....71
7.8. Conclusions....74

Classification of metabolites require the assignment of metabolites in NMR spectrum by experts or automatically using SSL, nevertheless, a more challenging situation is the detection of metabolites for which limited, or no spectral information is available in the training dataset. There is an emerging need for ND (novelty detection) when a class or classes are missing, poorly sampled or defined [80]. Basically, supervised, or semi-supervised training models enable only the prediction of metabolites which exist in the training dataset, whereas new or unexpected metabolites will be misclassified as an existing known metabolite. Applying ND is essential in metabolic profiling due to the complex nature of biological fluids and tissues. Metabolic variations in fluids and tissues can occur with any new stimuli and will cause alteration in the NMR spectra and new metabolites can appear in the NMR measurement. Therefore, using supervised or semi-supervised approaches might be insufficient in complex and – high-throughput NMR experiments. ND approaches are used to detect well-known and trivial components and discriminate potential new metabolites. These new metabolites are returned as candidates of new metabolites to the expert to manually assign them. Normally, ND is required in two situations. The first is when there are few examples to represent a known class within the training dataset; for instance, a particular category happens rarely, so the classification system does not have enough instances to represent this category. In this case, it is better to consider the rare category as novel or abnormal and test it against the model of normality. The second situation occurs when the training list is incomplete. Although enough instances are available to form a training model, it is expected that new classes will appear in the future [39]. In this chapter, we introduce the concept of ND of metabolites in 2D NMR TOCSY spectra where new metabolites are detected and assigned in a crowded spectrum using only the horizontal and vertical frequencies of the 2D TOCSY spectra.

Figure 7.1 summarizes the ND protocol: automatic peak picking is performed on the first 2D TOCSY spectra, two characteristic frequencies (F2, F1) are assigned to form the training dataset. The training models will be created based on the KNFST, SVDD and KDE classifiers with different training data volumes, observing the classifier performance and the corresponding execution time. The training model will be used in the testing phase to detect novel classes, i.e., novel metabolites in this case. Subsequently, the automatically derived peak picking parameters from the training phase are applied to the second TOCSY. The characteristic frequencies (F2, F1) are studied using the classifiers to identify novel peaks (i.e., metabolites) compared to the reference training models from the previous step. During the testing phase, training models are deployed to assess the novelty of particular metabolites and the success of the learning paradigm [252].This chapter has been adapted and/or adopted from [252].

7.1. KERNEL NULL FOLEY-SAMMON TRANSFORM

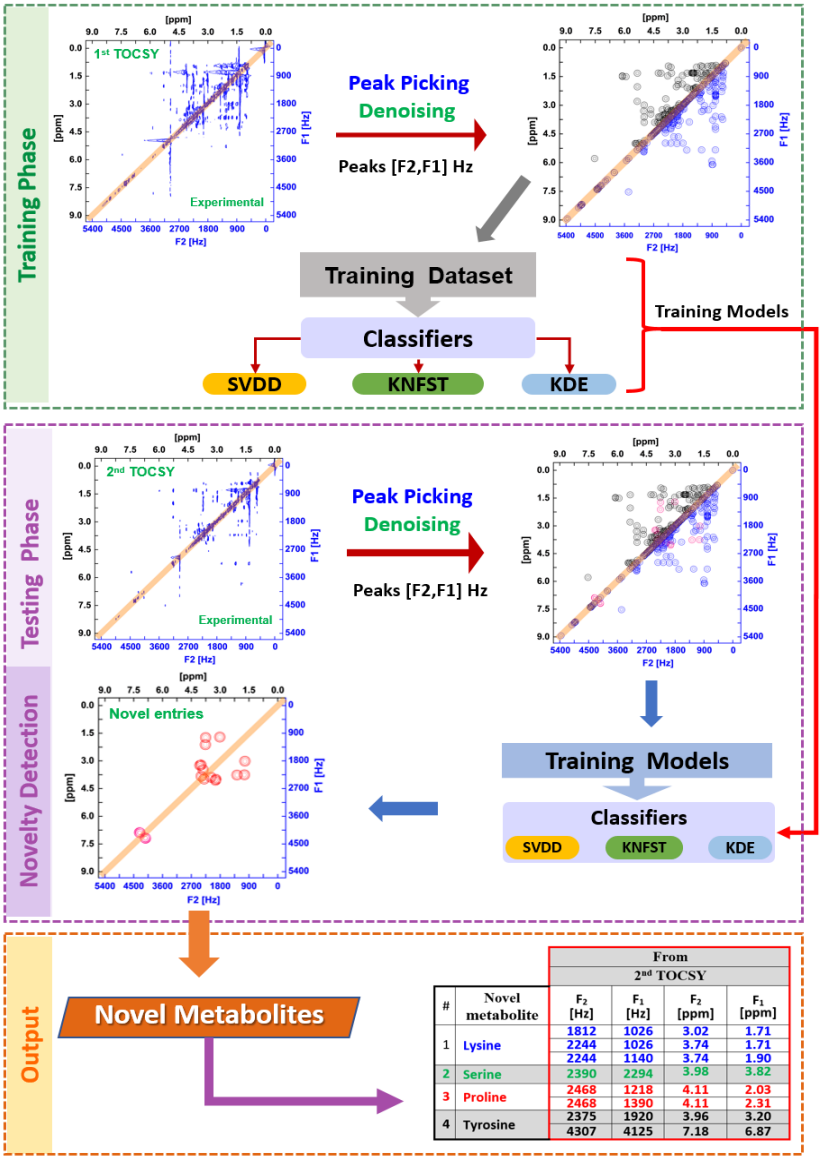
The Kernel Null Foley-Sammon transform (KNFST) was introduced under the SSL scenario using the confidence bands as an uncertainty measure in Section 6.3. In this section, KNFST is tested under the ND scenario. Similar to SSL, based on Eq. (6.14) and Eq. (6.15), ND KNFST tries to find the null projection direction matrix ω through minimizing the within-class scatter and maximizing the between-class scatter [241, 253]. KNFST is a joint multi-class model, which can achieve classification of all classes at once. The output of KNFST is used as a novelty score, where the larger the novelty score, the more novel is the test sample. A threshold is set to detect novelty borders. KNFST has been used in image classification [241, 253], gesture recognition [254], abnormal event detection in object tracking [255], authentication on mobile devices [256] and fault detection in machinery [257]. In this work, the KNFST code implementation in [241] has been customized.

7.2. SUPPORT VECTOR DATA DESCRIPTION

Support Vector Data Description (SVDD) is a domain-based method, which employs a hyperplane to represent a boundary based on the training data. This hyperplane tries to maximize the separation between different classes. SVDD was developed by [99] as a one-class classifier that distinguishes a positive (normal) class from all other classes in the dataset and builds its model based on the single positive class [80]. This approach creates a minimum-volume spherically shaped region that encompasses all or most of the training data of a chosen class. The hypersphere acts as a descriptor of normality, and a sample is considered an outlier if it falls outside the sphere [80, 258]. The problem of SVDD is an optimization problem that finds the center a with minimum radius R of the hypersphere that encloses most of the training data. SVDD enables the existence of outliers outside of the hypersphere, but a larger distance from the hypersphere is penalized in

$$\min_{R\in\mathbb{R},\xi\in\mathbb{R}^+} R^2 + C\sum_{i=1}^{n}\xi_i \qquad (7.1)$$
$$\text{subject to } \|\varphi(\mathrm{x_i}) - a\| \le R^2 + \xi_i$$

ξ_i is a slack variable that permits the existence of outliers, C is a parameter that controls the trade-off between the volume of the radius and the number of outliers (set to 1% in the thesis), and $\varphi(\mathrm{x_i})$ is the high dimensional mapping of x_i [99].



#	Novel metabolite	From 2nd TOCSY F_2 [Hz]	F_1 [Hz]	F_2 [ppm]	F_1 [ppm]
1	Lysine	1812	1026	3.02	1.71
		2244	1026	3.74	1.71
		2244	1140	3.74	1.90
2	Serine	2390	2294	3.98	3.82
3	Proline	2468	1218	4.11	2.03
		2468	1390	4.11	2.31
4	Tyrosine	2375	1920	3.96	3.20
		4307	4125	7.18	6.87

Figure 7.1: Schematic illustration of the ND procedure in metabolic profiling in a biological sample based on 2D TOCSY NMR spectra.

In this work, the binary classification implemented in the Novelty Detection Toolbox (NDtool) [38, 259] is extended to a multi-class approach using one-*vs*-all classification. SVDD has several applications in image and gesture classification [260-264], biomarker detection in HSQC NMR spectroscopy [265], and fault detection [266, 267].The novelty threshold of SVDD is defined as the radius of the hypersphere according to [99].

7.3. KERNEL DENSITY ESTIMATION

Kernel density estimation (KDE) is a probability-based method which computes the probability at each point in the data space within a localized neighborhood area of that point. KDE is a non-parametric approach that tries to estimate the probability of unknown distributions. The main assumption of density estimation is that samples reside in low-density areas indicate a low probability of being a known class. Accordingly, this area tends to contain novel data; whereas areas of high probability means the existence of known samples [38]. The probability density function is approximated by estimating the probability density through locating kernels at each point of the dataset, i.e., a kernel is centered at each data point, and then these kernels are summed up. A typical kernel density estimation is the Parzen Window estimator [34]. The Parzen estimator defines a fixed-width region $\Re$ centered at the sample point x and counts the number of neighboring sample points which falls in this region. Parzen estimators can be defined as:

$$p(x_i) = \frac{1}{N}\sum_{i=1}^{N} k_h\,(kx_j - x_i) \tag{7.2}$$

where $x_i \in X = \{x_1 \ldots x_n\}$, N is the number of data samples and kx_j are the region centers which are sampled from X. The density of x_i is calculated based upon the distance between kx_j and x_i and then representing it as a linear combination of the neighboring kernel centers. k_h is a kernel function centered at kx_j and has an associated parameter h related to the bandwidth parameter of region $\Re$ [268]. The parameter h is the Parzen window width. The Parzen width parameter is defined as the mean value of the distances between each kx_j and its k nearest neighbours. Since the probability must sum up to 1, we normalize the density by $\frac{1}{N_c}$ where N_c is the number of data points that belong to class c [34, 85]. KDE has been employed in tissue segmentation [269, 270], Alzheimer's disease detection in MRI [271, 272] and CT images [273, 274]. In this work, the binary classification implementation in NDtool [38, 84] has been extended to a multi-class approach using one-*vs*-all classification.

7.4. THRESHOLD SETTING AND NOVELTY DETECTION

Classifiers are designed to assign already known classes and, consequently, match the novel data sample to one of the known classes. ND tries to learn a model of normality, which is described by a novelty boundary. Normal instances are expected to be included in the normality model and reside within the novelty boundary, whereas unknown instances are expected to lie outside these boundaries [275]. A validation dataset is used to compute the novelty threshold for each known class in advance by finding the threshold with the minimum error on a validation dataset using grid search. During the testing phase, when classifying a data point, the threshold is compared to the output of the corresponding classifier. If the output does not comply with the pre-computed threshold, the data sample will be classified as novel. KNFST is a distance-based approach, which uses the assumption that similar data are located near each other, while novel instances are located away from known data. Thus, if the distance between the tested samples $d(z)$

is larger than the novelty threshold $\mathcal{T}$ of the class, the test sample is classified as novel, i.e., $d(z) > \mathcal{T} \rightarrow novel$. This is also valid for SVDD, where the radius of the hypersphere indicates the threshold. For KDE, if the posterior probability $p(x)$ is below the threshold $\mathcal{T}$, the more probable the test sample is a novel instance, i. e. $p(x) < \mathcal{T} \rightarrow novel$ [275, 276].

7.5. NOVELTY DETECTION OF METABOLITES USING BREAST CANCER TISSUE

The classifiers KNFST, SVDD and KDE are customized and tested for novelty detection of a breast cancer sample. The training data is partitioned into eight portions. These portions are used to test the system using different percentages of training data to observe the relation between the performance and the availability of training data and to examine the minimum size of the training set sufficient to yield a satisfactory performance. The portion of labeled training samples is increased every 50 cycles until all training samples are used in the classification process. In each cycle, different random permutations of training data are applied. The introduction of multiple cycles is vital; this is due to the random selection of the training data before starting the recognition process, which leads to different results for each chosen training dataset. Training portions of sizes 2.5%, 5%, 7.5%, 10%, 25%, 50%, 75% and 100% of the total training dataset size were used. In this experiment, a TOCSY spectrum of a breast cancer tissue sample, which comprises the metabolites in Section 5.3 is used.

To test ND on the TOCSY spectrum of breast cancer tissue, two scenarios are applied. The first scenario handles the one-class ND case. This experiment is built by excluding one of the metabolites from the training dataset, and afterwards a training model is built based on the remaining 26 metabolites. The testing dataset includes all 27 metabolites, which are the known 26 metabolites plus the excluded metabolite. On the second experiment, multi-class ND is employed by excluding multiple metabolites from the training set, and a training model is built based on the remaining metabolites. Subsequently, during the test phase the novelty scenario is tested based on the known and the excluded metabolites. In both scenarios, the classifiers are expected to detect the excluded metabolites and regard them as novel metabolites. The procedure is illustrated in Figure 7.2.

The assessment of the results is based on the ND metrics used in [277]. Let N be the total number of metabolites in the test dataset and N_c the number of novel metabolites in the test dataset.

- $M_{new} = (100 * F_n)/N_c$. The percentage of novel metabolites misclassified as known. F_n stands for the number of novel metabolites misclassified as known (*i.e., false negatives*).
- $F_{new} = (100 * F_p)/(N - N_c)$. The percentage of existing instances falsely misclassified as novel. F_p stands for the number of known metabolites misclassified as novel (*i.e., false positives).*

- $Err = 100 * (F_n + F_p + F_e)/N$. The percentage of total error Err where F_e denotes the misclassifications within known metabolites. It can be seen the Err includes also M_{new} and F_{new}.

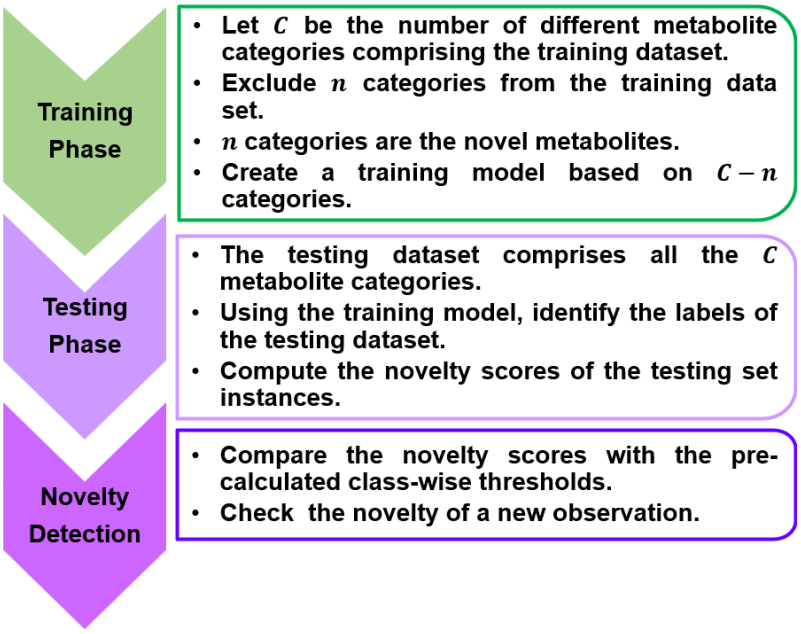


Figure 7.2: ND procedure by excluding one- and multi-metabolites from the pre-assigned 27 metabolites of the breast cancer tissue cell.

7.6. ONE-CLASS NOVELTY DETECTION

In the scenario of one-class novelty detection, the metabolite entry (tyrosine) is considered novel by excluding it from the list of 27 metabolites. Consequently, the training dataset consists of the remaining metabolites whereas the testing dataset includes the excluded novel metabolite tyrosine in addition to the known training data. Excluding a metabolite during the training process simulates the novelty of the excluded metabolite and ascertains that the training model is only aware of all metabolites excluding the exempted tyrosine. In breast cancer, tyrosine is the most frequent reported metabolic biomarker [278]. Figure 7.3(a-c) show the results of the ND procedure of the classifiers using the above assessment matrices for the metabolite tyrosine. Figure 7.3a shows that KNFST has a zero M_{new} rate regardless of the size of the training dataset, which means that tyrosine was correctly identified as novel. However, when using 2.5% of training data, in addition to misclassifying some known classes as novel classes, misclassification between known classes has a median error of 4%. On the other hand, using 2.5% of the training dataset, KDE and SVDD (Figure 7.3b and 4c) have a M_{new} value of around 4% and 50%, respectively, with a relatively high standard deviation. Both classifiers show zero M_{new} values after using only 5% of the training dataset. In general, for all classifiers the values of F_{new} and Err decrease when increasing the size of training samples. All classifiers achieve zero or near-zero values for M_{new} , F_{new} and Err when using 5% of the complete training dataset.

To test the overall performance of the system for all possible threshold settings, we use Receiver Operating Characteristic (ROC) curve analysis to show the tradeoff between false positives and true positives. ROC curves and Area under Curve (AUC) provide an assessment of the classification performance without indicating a decision threshold [110]. Figure 7.4 shows ROC curves which are generated using the one-*vs*-all approach for one run. This involves training one class per classifier, considering samples that belong to this particular class as normal samples and all other samples as novel [279].
As mentioned earlier, training portions of sizes 2.5%, 5%, 7.5%, 10%, 25%, 50%, 75% and 100% of the total training dataset size were used, nevertheless, for clarity only portions of sizes 2.5%, 10%, 100% are shown in the ROC curves, novelty scores and thresholds figures. These percentages give an indication of performance using relatively small, medium, and large amounts of training data. In general, it can be seen in Figure 7.4(a-c) that the classifiers' capability to distinguish novel metabolites from known metabolites increases by increasing the size of the training dataset. This can also be observed by the increasing values of the AUC, which implies a high diagnostic accuracy for large training data set sizes. Furthermore, it can be deduced that using 2.5% of the training data results in an inaccurate threshold, and consequently in a low recognition rate. By using 10% of the total training samples, the AUC of ROC curve of the metabolite tyrosine was over 97% for all classifiers. The AUC of the ROC curves are close to 100% for the three classifiers when using 100% of the training data.
Figure 7.5 shows the corresponding difference in novelty scores between known and unknown metabolites related to Figure 7.4. The red, green, and blue crosses resemble the unknown test data, known test data and known training data, respectively. The separation between the known and the unknown instances becomes more representative by increasing the training data size. In ideal cases, scores of known classes in the training dataset and testing dataset are similar. On the other hand, scores of novel instances must be relatively different.

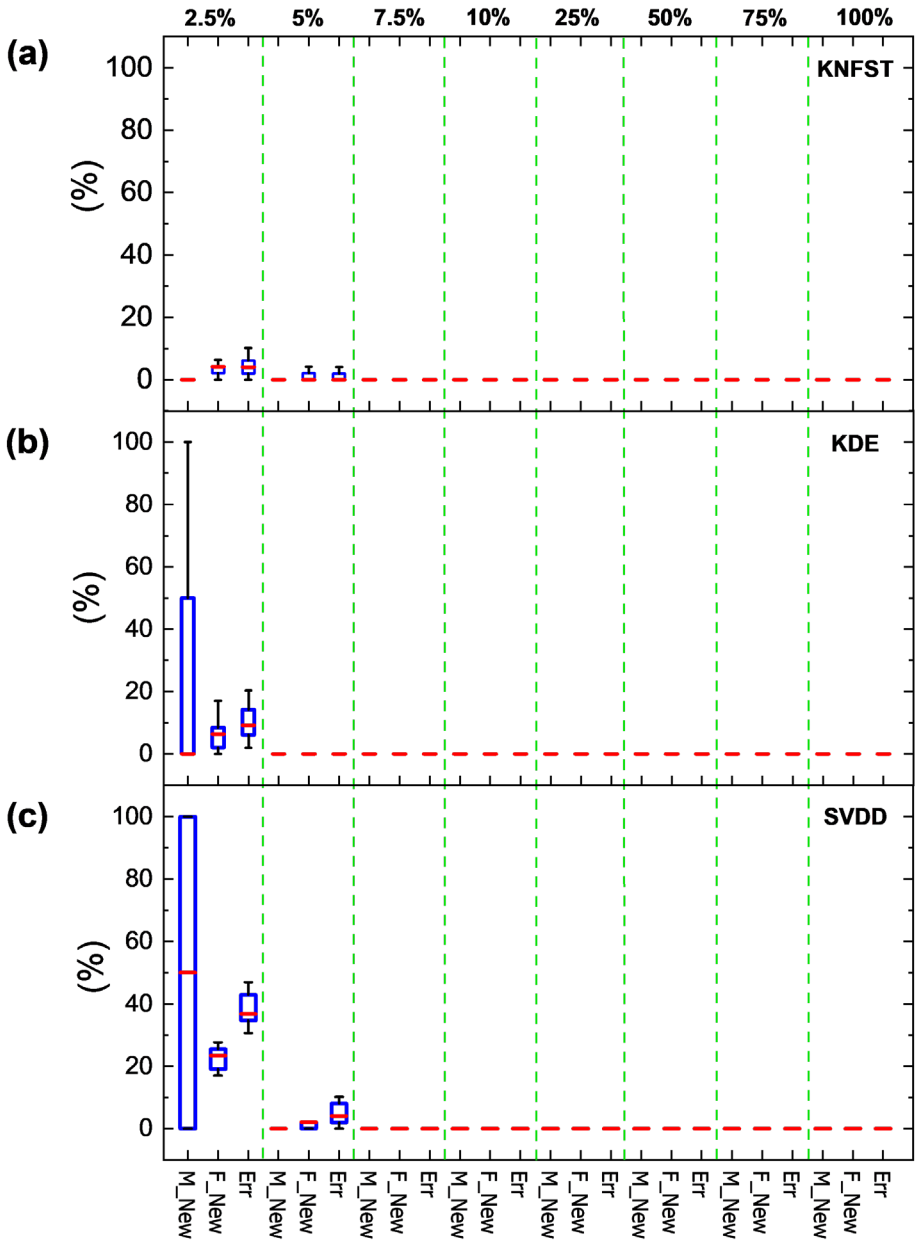


Figure 7.3: The M_{new}, F_{new} and Err values of breast cancer-tissue sample for the classifiers (a) KNFST, (b) KDE and (c) SVDD by applying one-class novelty detection.

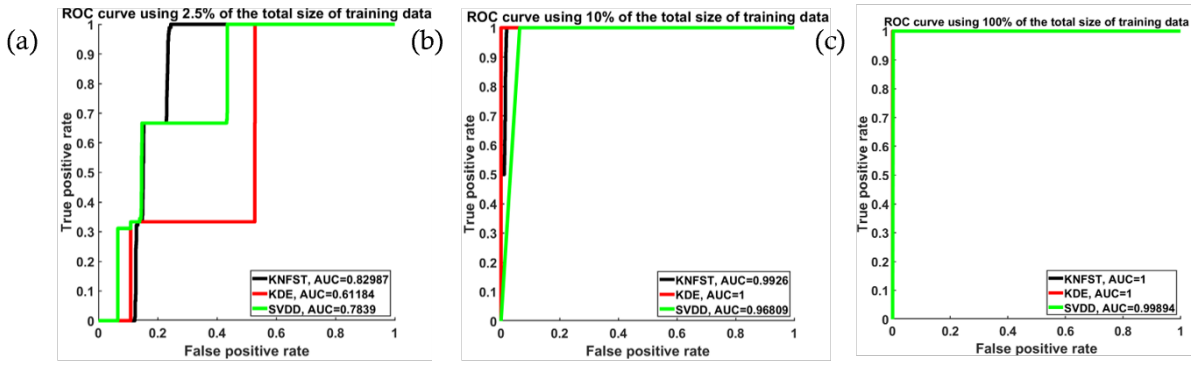


Figure 7.4: ROC curves and AUC values showing the accuracy of the novelty threshold for different sizes of training data for the metabolite tyrosine. From left to right, the ROC curve obtained using (a) 2.5%, (b) 10% (b) and (c) 100% of the total training dataset is shown.

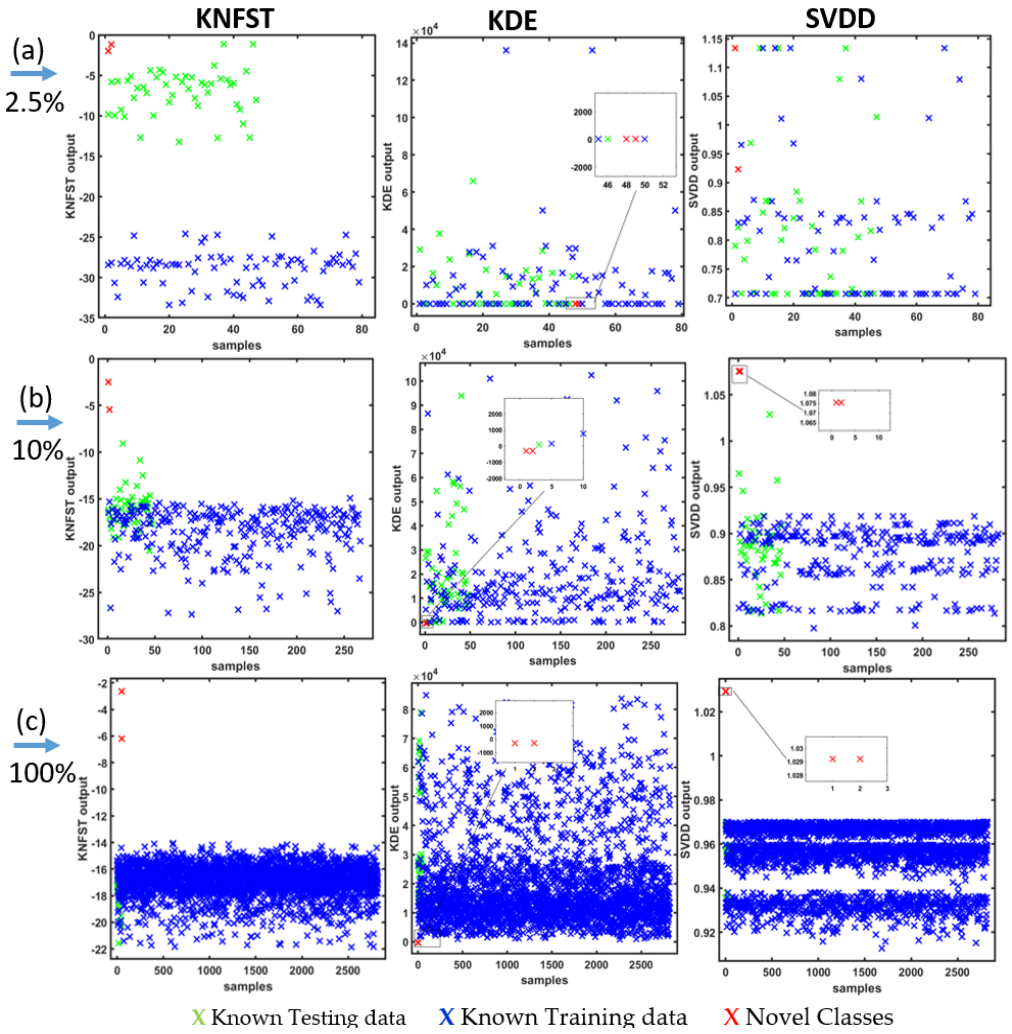


Figure 7.5: Novelty scores and threshold values of KNFST, KDE and SVDD classifiers using different training dataset sizes in the one-class novelty detection. The red, green, and blue crosses resemble the unknown test data, known test data and known training data, respectively. Subfigures (a) to (c) correspond to the variations of the output of the classifiers when using (a) 2.5%, (b) 10% and (c) 100% of the training dataset.

7.7. MULTI-CLASS NOVELTY DETECTION

According to [278], metabolites (leucine, tyrosine, proline and serine) are a subset of the clinically most frequently reported metabolic biomarkers related to breast cancer. Therefore, in the multi-class ND the above-mentioned metabolites were chosen to be excluded for novelty testing under different conditions. Accordingly, the classifiers were trained on 23 metabolites only. During the test phase, all assigned 27 metabolites of the breast cancer sample were included in the test dataset, likewise the one-class novelty detection.

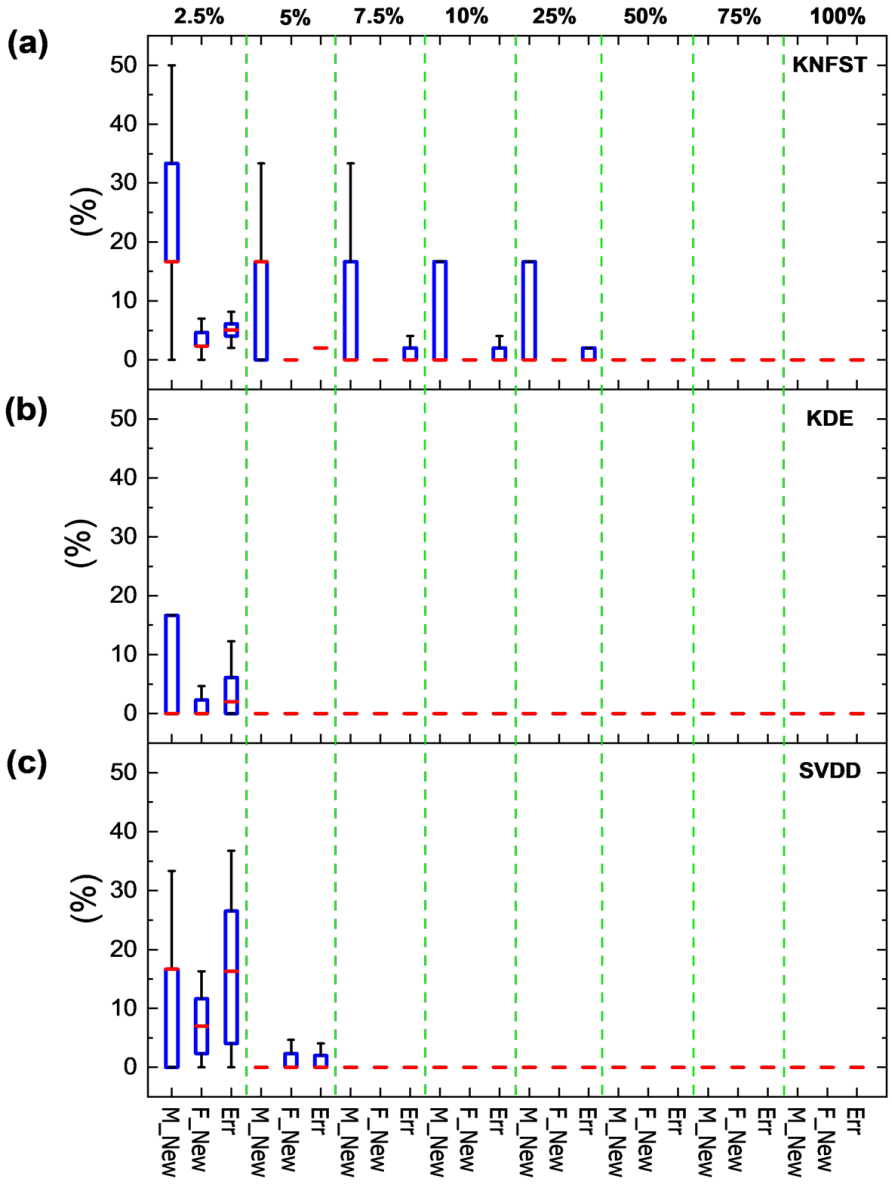


Figure 7.6: M_{new}, F_{new} and Err values of breast cancer tissue sample for the classifiers (a) KNFST, (b) KDE and (c) SVDD by applying multi-class novelty detection.

Figure 7.6 shows the M_{new}, F_{new} and Err values in multi-class ND scenario. When using 2.5% of the training data, KNFST and SVDD have similar M_{new} median values around

16%. The SVDD M_{new} distribution shows a negative skewness, which means most M_{new} values are low. Although KDE has a median of zero M_{new}, KDE and the other classifiers have a high standard deviation. This means a low discrimination capability at extremely low training dataset size. Similarly, the values of F_{new} and Err showed unstable standard deviations and median values in all classifiers. Starting from 5% training data size, KNFST showed a negative skewness in M_{new} values, which implies a progressing discrimination of novel metabolites. On the other hand, KDE and SVDD have zero for M_{new} and approximately zero value for F_{new} and Err. Starting from 50% of the training data size, a median of zero M_{new} values were reached for KNFST. Using only 25% of the training data, all the classifiers have reached less than 3% median values for M_{new}, F_{new} and Err values. In addition, already when using only 5% of the training data, all classifiers reached near-zero median values of F_{new} and Err, indicating that the classifiers are able to correctly classify known metabolites and detect novel instances.

Figure **7.7** (a-c) shows novelty scores of the KNFST, KDE and SVDD classifiers using 2.5%, 10%, and 100% training dataset size by applying the multi-class novelty detection. The red crosses correspond to the six-pattern related to tyrosine, proline, leucine, and serine. Comparable to one-class novelty detection, the novelty threshold becomes more accurate and the separation between normal and abnormal instances becomes more distinct when increasing the training dataset size. Remarkably, an acceptable threshold could be calculated even when only 10% of the training data were considered.
Unlike one-class classification, generating ROC curves for multi-class classification tasks is not a straightforward solvable problem. A typical solution is to generate individual ROC curves for each class separately using the one-vs-all method [110]. Figure 7.8 shows the mean and standard deviation of the total classification processing time of 50 runs in the one- and multi-class novelty detection. The experiments were run on Windows 10 using an Intel Xeon E5 machine with 16 GB memory and 2.8 GHz Quad Core CPU. The computational complexity for KDE is $O(N^2)$ [280], and $O(N^3)$ for KNFST [241] and SVDD [281]. The execution time for KNFST and SVDD grows when increasing the amount of training data. In one-class novelty detection, the execution time for KNFST increases steadily until it exceeds the SVDD execution time. However, rather than increasing, the execution time for one- and multi-class novelty in KDE remains almost constant when increasing the size of the training dataset. This might be due to the fixed Parzen window width of the kernel used by KDE. The estimation of the optimal Parzen window width is the most effecting computational factor [280]. As stated earlier, the Parzen width parameter is defined as the mean distance between the k-nearest neighbors and the instances in the training dataset. The number k of neighbors in our experiments was two [282]. In SVDD, computational cost is related to tuning the parameters of the kernel, and there is a direct relation between the size of the training dataset and the execution time [283]. This can be seen SVDD time consumption on multi-class ND where, in comparison to the one-class scenario, more novel samples are encountered. The main computational cost in KNFST comes from computing a joint kernel feature space for all known classes and the eigenvalue decomposition of the kernel matrix [241, 284].

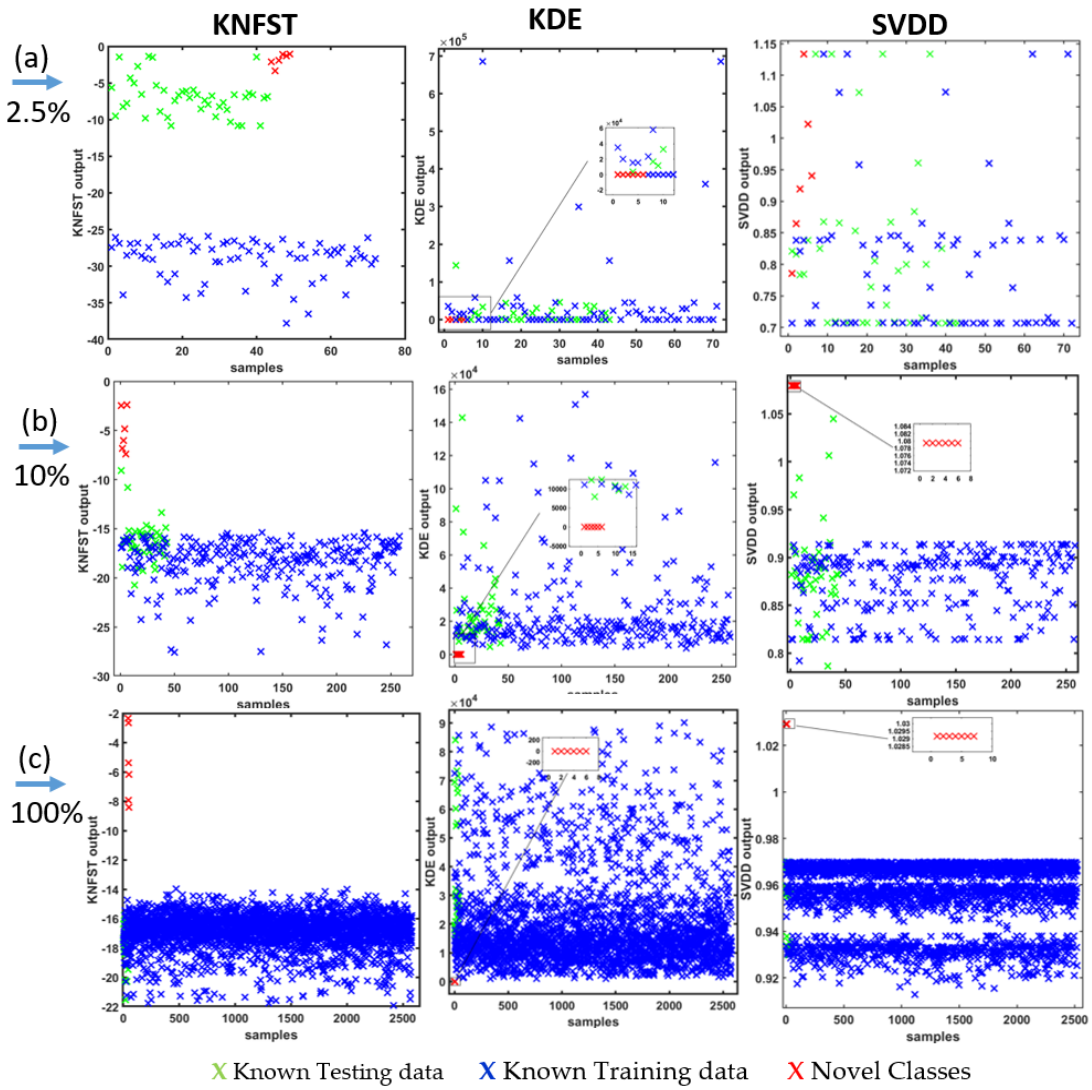


Figure 7.7: Novelty scores and threshold values of KNFST, KDE and SVDD classifiers for different training data size for multi-class novelty detection.The red, green, and blue crosses resemble the unknown test data, known test data and known training data, respectively. The output of the classifiers is shown for (a) 2.5%, (b) 10% and (c) 100% of the training dataset.

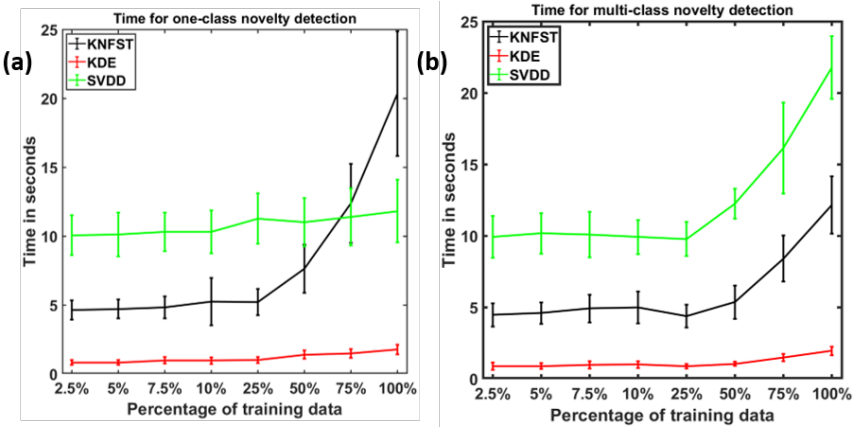


Figure 7.8: Total time from training to classification for (a) one-class and (b) multi-class novelty detection.

The confusion matrices of one- and multi-class novelty, in addition to the ROC curves for the multi-class ND algorithm, are presented in the appendix A.

7.8. CONCLUSIONS

In this work, ND was established based on 2D NMR TOCSY spectra for metabolic profiling associated to dynamics changes in biological systems, where metabolites of breast cancer tissue samples were extracted from the TOCSY spectrum. The one- and multi-class ND tests were designed to consider peak assignments appearing in the TOCSY spectrum as a reference database. Subsequently, one and four metabolites were excluded from the reference TOCSY to simulate their novelty. The KNFST, KDE and SVDD classifiers were tested to detect the excluded metabolites. The classifiers achieved explicit labelling to metabolites that appear in the TOCSY and additionally detected new metabolites which are unknown to the training model. Despite the observed overlapping in the training dataset resulting from chemical shifts, the implemented methods in this work achieved 0% false positive rates at 100% true positive rate. The resulting classification performance increases with increasing training dataset size. Generally, the execution time also increases when increasing the training dataset size for all classifiers, nevertheless, the execution time is noticeably short. The results are supported by confusion matrices and ROC curves in addition to plotting the novelty outputs. The presented machine learning based ND techniques provide promising perspectives for automated assignment of metabolites that evolve in dynamic biological environments and trigger the metabolic pathways.

8. Contribution: Automated Monitoring of Metabolic Changes Accompanying the Differentiation of Adipose Tissue-Derived Human Mesenchymal Stem Employing ^{1}H-^{1}H TOCSY NMR

8.1. Machine learning 76
8.2. Metabolic evolution of AT-derived hMSCs 76
8.3. Conclusion 82

Mesenchymal stem cells (MSCs) are multipotent stem cells with high capacity to proliferate and differentiate, while exhibiting low immunogenicity and providing immunosuppressive properties [285]. These potentials put MSCs in the lead as a promising candidate for several innovative strategies of cellular therapy and tissue engineering. MSCs are obtained from several body tissues, and their potential to reproduction and developmental is highly dependent on their source of origin [286]. Adipose tissue is considered a highly valued source to isolate MSCs being a byproduct that generate a high yield of primary cells, with high potential to proliferate and differentiate; therefore, adipose tissue-derived MSCs are applied highly in tissue engineering and regenerative medicine [287]. Metabolic adaptation of MSCs is highly dependent on their surrounding environment; MSCs cultivated under hypoxic condition show limited proliferation rate and high production of glycolytic enzymes, while in normoxic conditions they show high proliferation rate and an additional reliance on oxidation phosphorylation aside with glycolysis, in what its named by Warburg effect [288]. In addition, the differentiation of MSC into adipocytes and osteocytes was shown to be accompanied by a high level of oxidative phosphorylation, in fact, studies have shown that the differentiation of MSCs into osteocytes is negatively affected under normoxic conditions [289]. The switch between the glycolytic and oxidative phosphorylation pathway shows the flexibility of MSCs in adapting a metabolism that enable them to fulfil their role at the site of their residency [290]. New approaches are required to reveal novel biomarkers and information in the metabolism of MSCs and to track the metabolism states in response to stimuli, and metabolic adaptation associated with several biological processes, including differentiation [197, 291]. This information may unveil their behavior to be controlled and guided toward successful therapies providing the proper culture conditions and handling [292].

In this chapter, machine learning is applied to automate the monitoring of the MSCs differentiation and to resolve the convolution of the associated NMR spectra using the approaches introduced on Chapter 7. Furthermore, through automating non-targeted metabolic profiling, the dynamic evolution of biological samples will have an unlimited

perspective and will overcome the inherent obstacles in non-targeted 2D NMR analysis. Figure 8.1 demonstrates the experimental settings followed in this chapter. AT-derived hMSCs are cultivated in a basal culture media and measured after four days using NMR. Non-targeted metabolic profiling of 2D NMR TOCSY is generated based on the four days cultivation where all collected peaks are manually assigned by the expert. AT-MSCs were subdivided into three experiments. On the first one, the MSCs were maintained in basal MSCs culture for prolonged cultivation. On the second and third experiments, AT-MSCs were induced to differentiate into adipocytes or osteocytes respectively. After fourteen days, the adipogenic and osteogenic differentiation of the AT-derived hMSCs in addition to their control group were measured using 2D NMR TOCSY. Similarly, peak-picking is applied, and the cross peaks are assigned by an expert. To evaluate the performance of our methodology, the manual assignments are compared by the automated method. This work was adopted/adapted from [293].

8.1. MACHINE LEARNING

To monitor the dynamic evolution of adipose tissue-derived human MSCs (AT-derived hMSCs) using 2D NMR TOCSY spectra, KNFST and KDE were used.

8.2. METABOLIC EVOLUTION OF AT-DERIVED HMSCS

To observe the dynamic evolution of the AT-derived hMSCs at after 14 days of cultivation (Ct d14) and 14 days of adiobocytes (AT d14) and osteocytes (OS d14) differentiation, the training dataset created from (Ct d4) is used to create the main training model $\theta_{CT\ d4}$ using KNFST and KDE. Three independent testing datasets are constructed using Ct d14, AT d14 and OS d14 using the corresponding frequencies in Table 5.2, and are introduced to the classifiers and tested against $\theta_{CT\ d4}$ The results are reported as multi-class confusion matrices that compare the human-based metabolic profiling described in Section 5.4.2 with the predicted assignments of the frequencies of the TOCSY spectra. In addition, Figure 8.5 shows the novelty scores produced by the classifiers to show the separation ability of the classifier in terms of projection distance for KNFST and probability estimation for KDE. The scores are color-coded to distinguish the scores of the different representations of classifier outputs as follows: the scores of known instances in the training set in blue, the scores of known instances in the testing dataset in green, the scores of missed novel instances in pink, the scores of correctly classified novel classes in red and the scores of misclassified known instances in the testing dataset in black. In ideal cases, the scores of known classes in the training dataset and testing dataset are similar. On the other hand, the scores of novel instances must be relatively different to those known classes. Novelty thresholds are created based on the validating dataset choosing the thresholds with a minimum validation error.

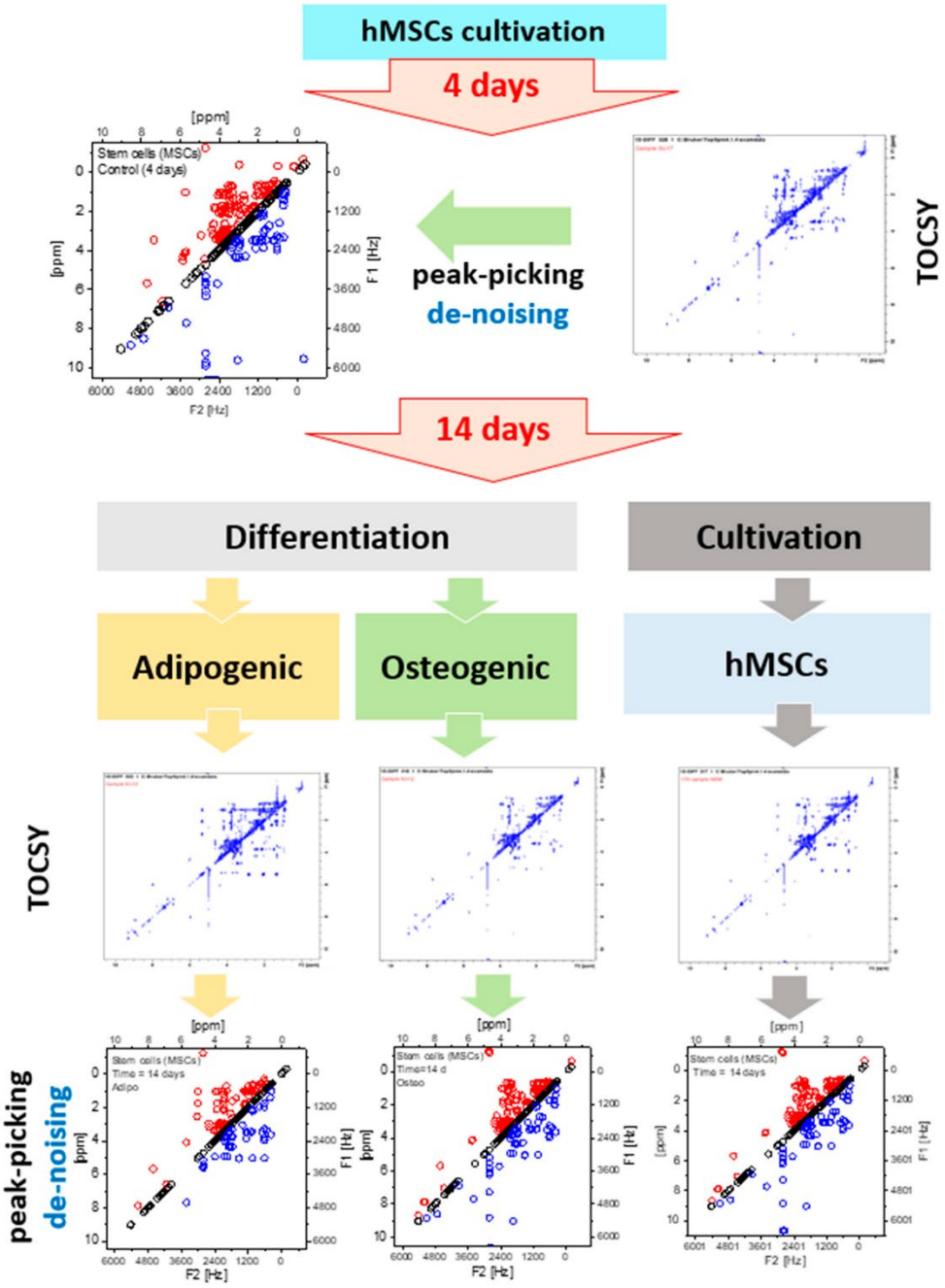


Figure 8.1: Schematic diagram of the experimental setting to observe the metabolic evolution of AT-derived hMSCs.using 2D TOCSY of intracellular extracts of MSCs cultivated in basal culture media at 4 and 14 days and MSCs cultivated for a duration of 14 days in an adipogenic and osteogenic differentiation media.

Ct d14: Figure 8.2 shows the confusion matrices for the output of the classifiers KNFST and KDE for Ct d14 sample. Both classifiers were able to detect all the sixteen novel frequencies which belong the fatty acids, 1-methylnicotinamide, myo-inositol, and taurine in the sample. No misclassification was encountered in KDE as observed on Figure 8.5b. Nevertheless, KNFST had two misclassifications within known classes, where the two instances of valine were misclassified as proline. This can be seen in Figure

8.5a, where two instances were plotted in pink, indicating the misclassification within known classes.

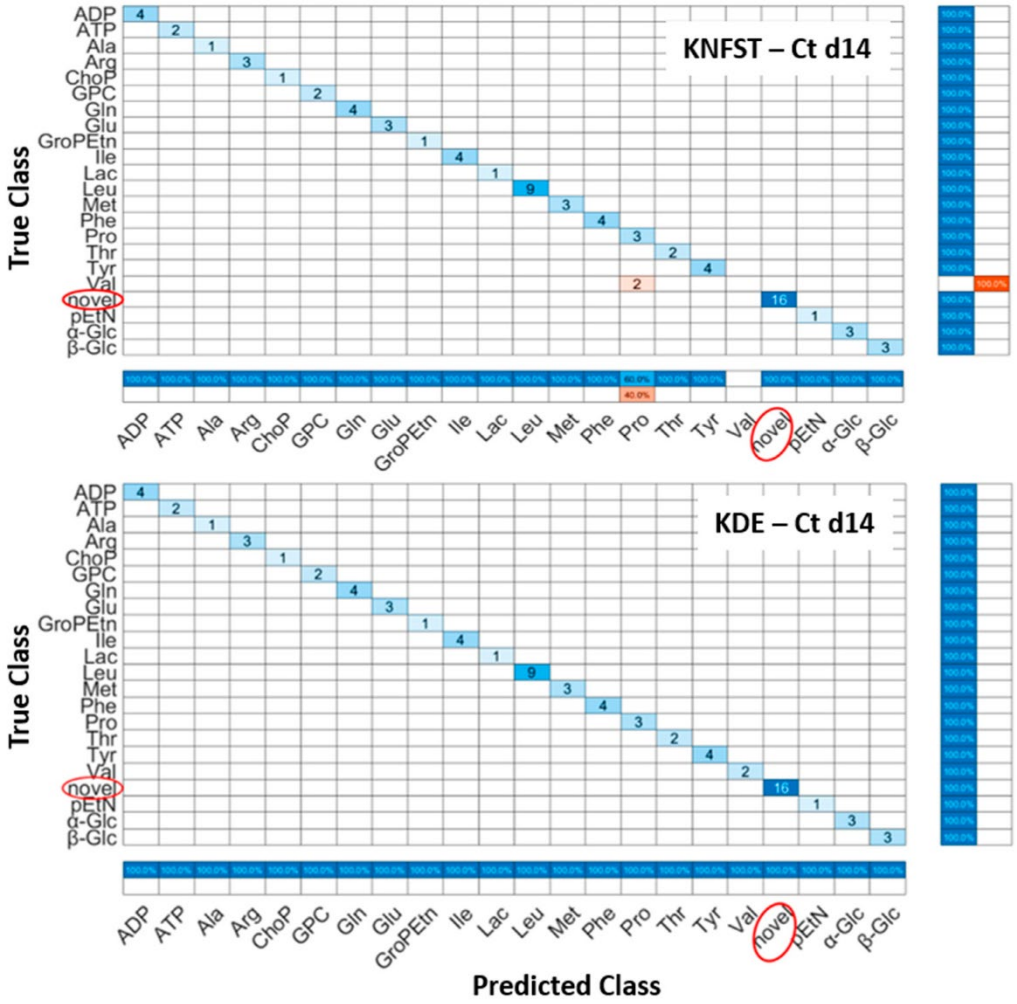


Figure 8.2: Confusion matrices of the output of classifiers KNFST and KDE for the spectrum after 14 days cultivation.

AT d14: It can be seen on Figure 8.3 that both classifiers predicted all the sixteen novel metabolites which belong to the fatty acids, 1-methylnicotinamide, myo-inositol, and taurine in the sample. Nevertheless, both classifiers had misclassification within already known classes. KNFST and KDE misclassified methionine as glutamine. In addition, KNFST misclassified one of the instances of valine and proline as well as misclassified one instance of leucine as threonine. This can also be seen on Figure 8.5c,d where misclassifications of known classes were plotted in pink.

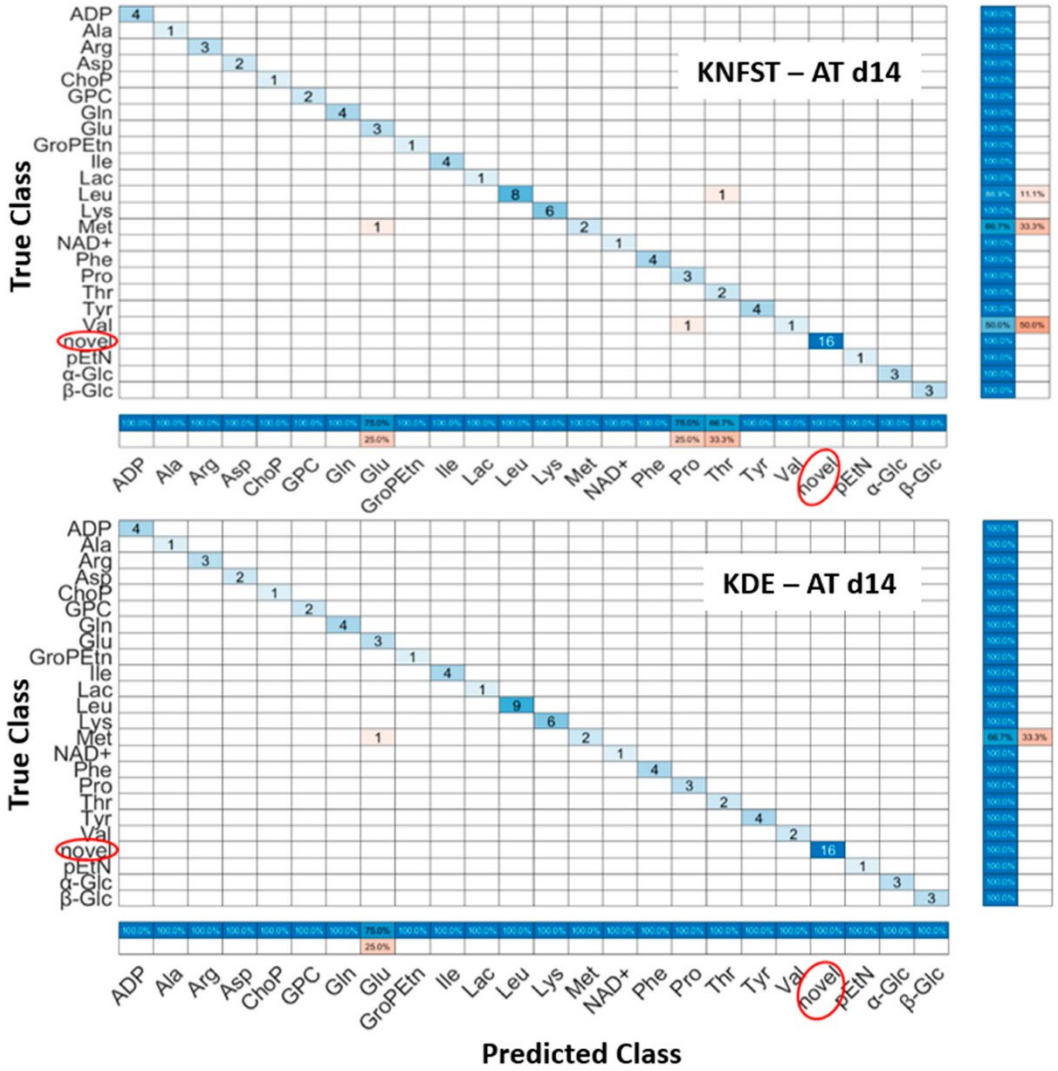


Figure 8.3: Confusion matrices of the output of classifiers KNFST and KDE for the spectrum of 14 days adiobocytes differentiation.

OS d14: Figure 8.4 shows the confusion matrices for the output of the classifier KNFST and KDE for the OS d14 sample. Both classifiers were able to detect all six novel instances in the sample, such as myo-inositol, Fat2 and taurine. However, it can be observed that valine was misclassified as proline in KDE. This may be due to the overlap in the vertical and horizontal frequencies between these metabolites, which can be seen in Table 5.2 and Figure 5.4d. Except for this single misclassification, no misclassification was encountered in both classifiers. This can be also observed in Figure 8.5e, f.

Depending on the test sample, the number and type of novel metabolites differ. For instance, there are 16 identical novel (but shifted in frequency) metabolites in Ct d14 and AT d14 in comparison to Ct d4. Nevertheless, the disappearance of metabolites in both samples is also different. In sample OS d14, six metabolites were found in comparison to Ct d4, and more metabolites disappeared during the differentiation. For both classifiers and all samples, the disappearance of metabolites during the biological pathway did not affect the classification performance. For instance, though the main training

model $\theta_{CT\ d4}$. was created on specific metabolites that disappeared in the spectra of Ct d14, AT d14 and OS d14, both classifiers proved their classification flexibility in observing metabolites presence and absence. Hence, the classifiers were able to detect both the presence and the absence of individual metabolites in accordance with $\theta_{CT\ d4}$.

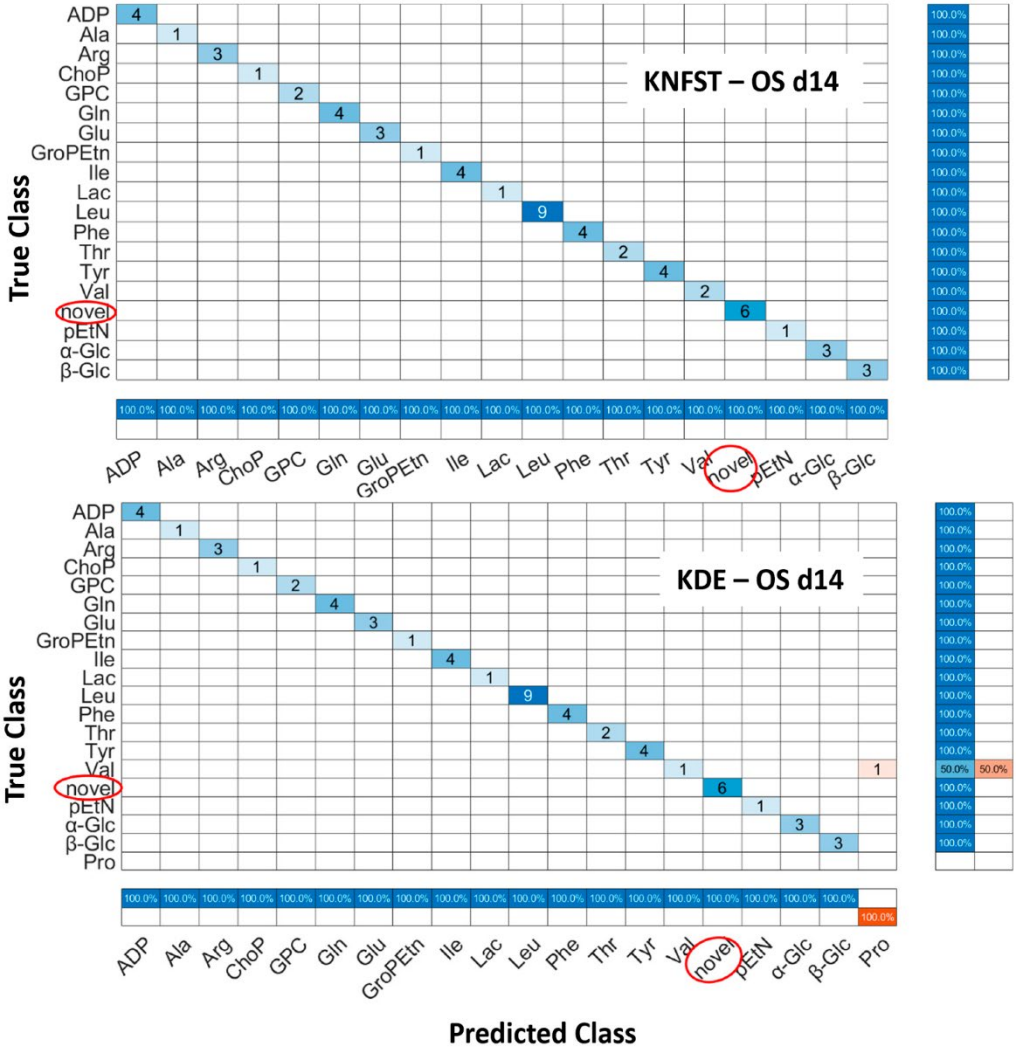


Figure 8.4: Confusion matrices of the output of classifiers KNFST and KDE for the spectrum of 14 days osteocytes differentiation.

Following the novelty detection metrics used in Section 7.5, Table 8.1 summaries the performance of the classifiers subject to the sample type.

Table 8.1: A summary of the performance of KDE and KNFST classifiers for Ct d14, AT d14 and OS d14.

	Ct d14		AT d14		OS d14	
	KNFST	KDE	KNFST	KDE	KNFST	KDE
False negative rate	0%	0%	0%	0%	0%	0%
False positive rate	0%	0%	0%	0%	0%	0%
Total error	2.6%	0%	3.6%	1.2%	0%	1.7%

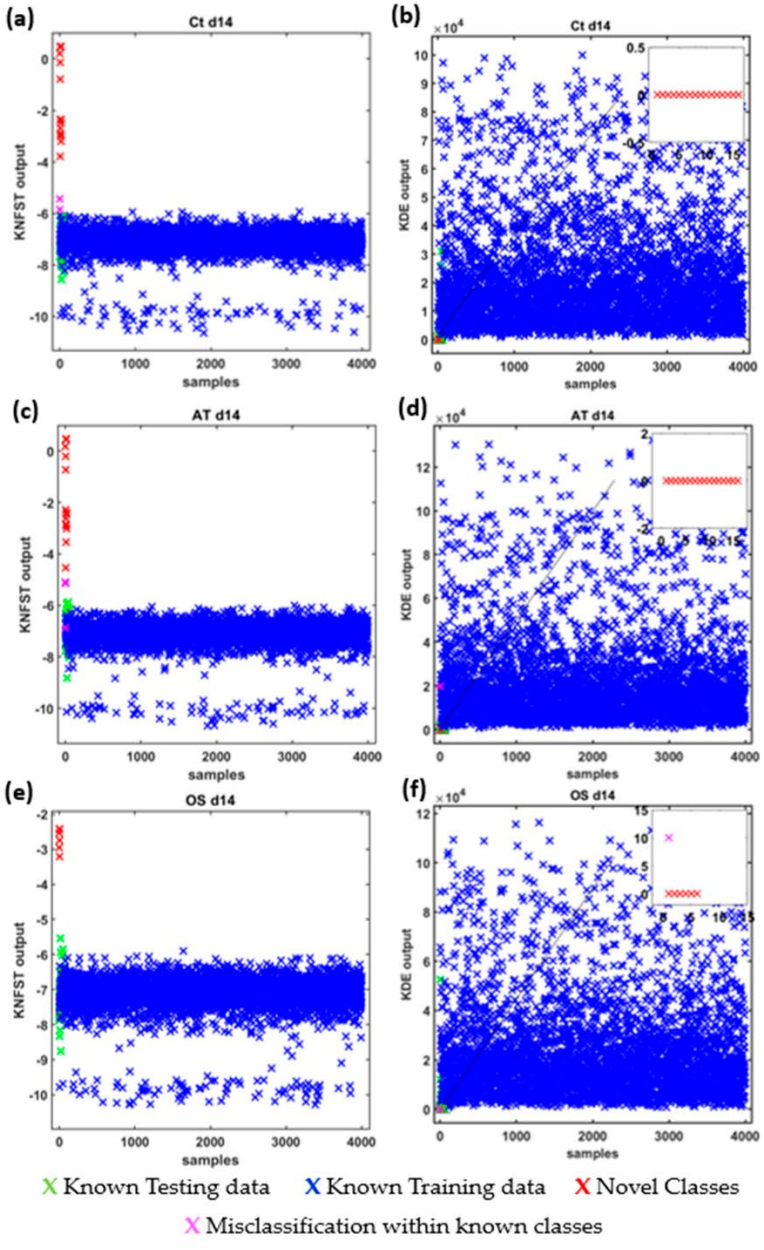


Figure 8.5: Novelty scores and threshold values of KNFST and KDE classifiers for Ct d14 (a, b), AT d14(c, d) and OS d14(e, f).

8.3. CONCLUSION

In this chapter, we demonstrate using machine learning to perform an automatic analysis of ^{1}H-^{1}H TOCSY spectra acquired on cultivated and differentiated adipose-tissue-derived human MSCs (AT-derived hMSCs). Multi-class classification in addition to the novelty detection of metabolites were established based on four different 2D NMR TOCSY spectra. The primary training model was built using TOCSY spectrum of AT-derived hMSCs at four days of cultivation. Subsequently, the metabolic changes of AT-derived hMSCs control sample were monitored under three different biological settings employing the classifiers KDE and KNFST. Despite the severe overlapping in the frequencies in TOCSY spectra, the classification outputs proved the efficiency of the used method. KDE and KNFST achieved a total classification error between 0% and 3.6% and false positive and false negative rates of 0%. The investigation in this work confirms the common metabolic pathways associated with stem cell biology. In the future, further features can be added to the dataset to produce a higher discriminative ability. Furthermore, chemical structure information or integrating other 2D NMR spectra can be included in the classification process. This work provides methodological approaches to track information of MSCs metabolism and their biological pathways, including detecting novel metabolites related to diverse stimuli in terms of prolonged cultivation and varied differentiation. This work can be extended to monitor further kinds of MSCs proliferation and recognize spectral signatures of pathways and processes.

9. Summary and Conclusions

Machine learning based methods are promising tools that can extract concealed knowledge from biological data. This information can be used to relate the data to dynamic modeling of biological systems to get an evident and improved comprehension of data and diseases. Recently, various methods that are related to automatic metabolic assignment in NMR have been proposed. The metabolic profiling concept of these methods is based on using deep learning to analyze contours images of the TOCSY spectrum or uses multivariate analysis techniques. In contrast to these approaches, the methods developed in this thesis are based on employing the frequencies of the TOCSY spectra in the analysis process. In NMR, frequencies operate as a metabolic fingerprint of potential biomarkers. The use of the horizontal and vertical frequencies is beneficial because frequencies are related to the standard ppm values of the chemical shifts of metabolites. Depending on the NMR spectrometer frequency, chemical shifts given in ppm and frequencies are easily exchangeable. Moreover, chemical shift frequencies are considered the most informative variable in NMR [145] and they can be consistently reproduced under pre-established protocols.

In this thesis, multiple machine learning methods have been proposed to enable automatic and accurate spectral assignment of metabolites based on deconvolution of 2D-TOCSY NMR spectra. Semi-supervised learning and ND techniques based on third- and fourth-degree polynomial classifiers, Kernel Null Foley-Sammon transform, Support Vector machines and Kernel Density Estimation are presented.

In Chapters 6 and 7, metabolic profiling associated to dynamic changes in biological systems were studied. One these Chapters, 27 metabolites from breast-cancer tissue samples were extracted from the 2D NMR TOCSY spectrum to be used in used in the automatic metabolic profiling experiments. Semi-supervised learning of 2D NMR is essential due to the spectral components induced by chemical shifts, overlapping of metabolites, noise, and biological matrix effects, which aggravate the metabolic annotation process even for experts. In addition, manual labeling is expensive in terms of time and effort and particularly dependent on the expert's experience. Confidence bands were used to accept or reject the classification results of semi-supervised learning. Based on our results, SSL can be used as a strong and confident replacement for the manual assignment of metabolites in 2D NMR spectra. Novelty detection is vital in metabolism due to the nature of biological systems where new metabolites can emerge because of dynamic interactions within cells or different stimuli that trigger change. In Chapter 7, one- and multi-class novelty detection experiments were employed. Subsequently, one and four metabolites were excluded from the reference TOCSY to simulate the novelty of the extracted metabolites. The performance of the algorithms has been evaluated according to different training data sizes through matching the results deduced by human specialists with the output of the novelty detection. The results have shown that despite the obvious overlapping, the implemented methods in this work achieved high performance and low mislabeling rates.

In Chapter 8, multi-class classification in addition to novelty detection of metabolites was established based on four different 2D NMR TOCSY spectra. The analysis is based on comparing the intracellular metabolites of the control cultivation on a basal culture media at four days and the successive metabolic evolutional on the same cell at fourteen days of cultivation in addition to their adipogenic or osteogenic differentiation for a duration of fourteen days. The classifiers Kernel Null Foley-Sammon Transform and Kernel Density Estimation achieved a total classification error between 0% and 3.6% and false positive and false negative rates of 0%. This approach was successfully able to automatically reveal metabolic changes that accompanied MSC cellular evolution starting from their undifferentiated status to their prolonged cultivation and differentiation into adipocytes and osteocytes using machine learning. The investigation in Chapter 8 strengthens the conclusion derived from Chapter 7, because it is consistent with the real metabolic pathways that are observable in stem cells research [197, 291, 292]. While in Chapter 7 a simulated novelty system has been tested, the study in Chapter 8 investigated a real and confirmed metabolic pathway that has been initiated through different biological triggers.

Future work

For future strategies, creating a more comprehensive and standardized metabolic database using ppm, horizontal and vertical frequencies designed for different NMR resolution frequency is essential to stimulate an uncomplicated access to diverse NMR data. This perspective is critical due to the heterogeneity of metabolites and the associated variables and implications. Furthermore, a new feature, which is related to spin–spin couplings, can be added to the two already existing features to increase the discriminative strength. Moreover, additional 2D NMR methods such as HMBC or HSQC can be employed and integrated in the automatic prediction. The output of the classification using different techniques might then be combined as ensemble classification to generate more accurate results in more complex mixtures. Quantification of the NMR signal is a planned goal for future developments. The introduction of the quantitative characterization in the classification process will result in a comprehensive and quantitative analysis of 2D NMR TOCSY.

The proposed methodologies are aimed to accelerate and facilitate the metabolic profiling of 2D NMR. This development is a big step forward in automated spectral assignment and a backbone for future enhancement and development in this area.

10. Appendix

A. Novelty detection related results

i) Confusion matrices and ROC curves

An important statement is the following: For programming indexing purposes, the metabolites are reordered considering the novel metabolites. This does not affect the frequencies and only serves as a programming maneuver. The novel metabolites are renamed and shifted to the last index, which explains the variations on the labels of the novel metabolites in Figure Supp. 1 to Figure Supp. 5 in comparison to the labels in Table 5.1. Example for one class-novelty: All classes: a, b, c, and d. The novel class: b. Reordering:

Index	1	2	3	4
Classes	a	b	c	d
Reorder/Rename	a	c	d	b

Therefore, class b is shifted to the end index. Example for multi-class novelty:
All classes: a, b, c, and d. The novel classes: a and b
Reordering:

Index	1	2	3	4
Classes	a	b	c	d
Reorder/Rename	c	d	b	a

The confusion matrix is utilized to describe the performance of the classification algorithm in terms of true positive, true negative, false positive and false negative values. Figure Supp. 1 to Figure Supp. 3 show the confusion matrices of one single run using different training dataset sizes for one-class novelty detection by excluding the metabolite tyrosine from the training dataset. It can be observed that, despite the variation in the training data size and the error in identifying known classes, the classifiers were always able to detect the novel metabolite, which is indicated as metabolite 27. A red flag on these figures indicates which class is novel. A row summary is included on each figure in cases where a severe misclassification is present. The results of Figure Supp. 1 to Figure Supp. 3 are consistent with the performance measures, Figure Supp. 4 to Figure Supp. 6 show the confusion matrices of one-run using different training dataset sizes for multi-class novelty detection by excluding metabolites leucine, tyrosine, proline, and serine from the training dataset. In comparison to the one-class novelty case, it can be observed that the classifiers were unable to detect all novel classes, indicated as metabolites 24, 25, 26, 27, using a small size of the training dataset. The detection of novel classes is improved for larger sizes of the training dataset. A red flag on these figures indicates the novel classes. The results of Figure Supp. 4 to Figure Supp. 6 are consistent with the performance measures and novelty scores figures in the main manuscript and the ROC curves in the supplemental material.

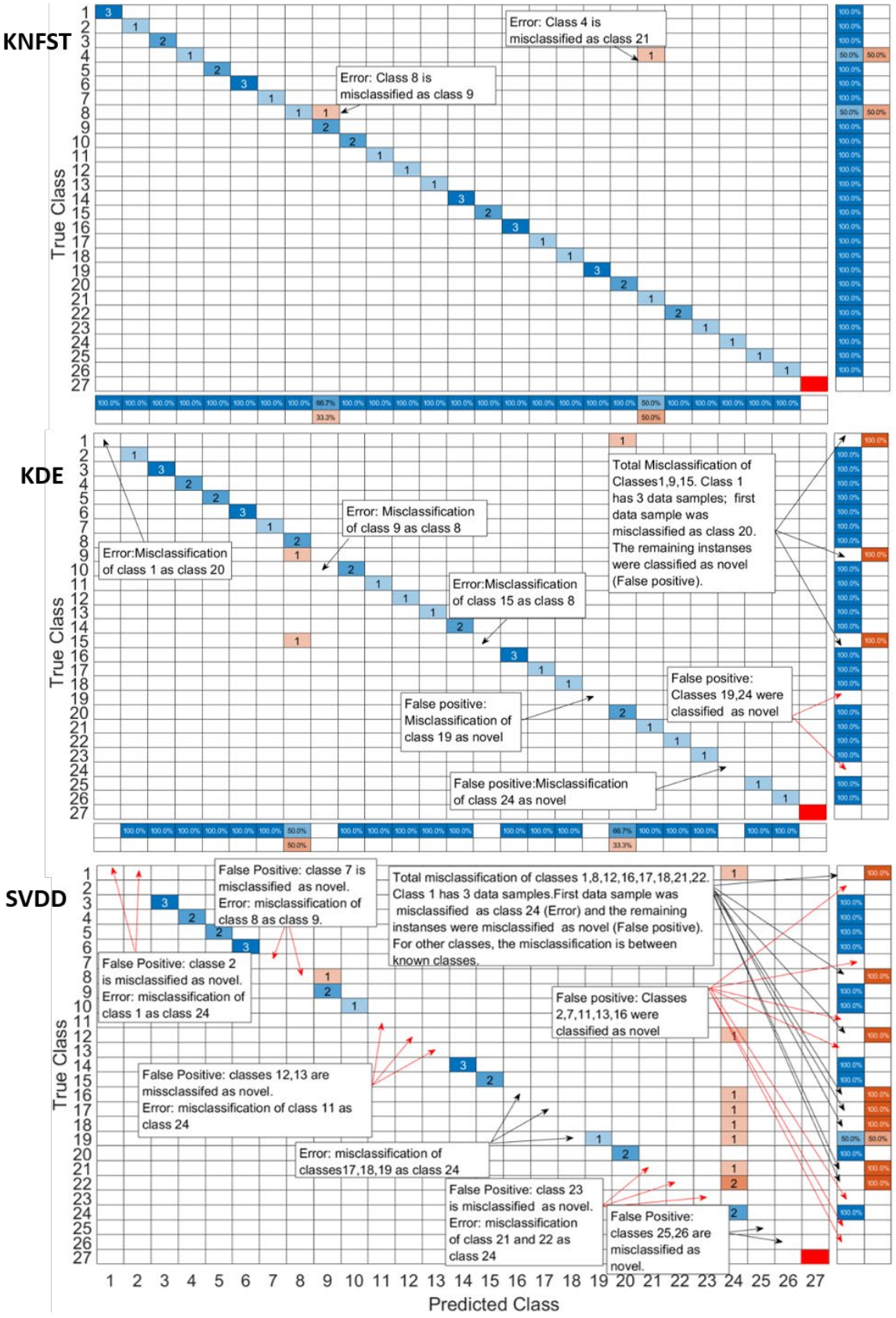


Figure Supp. 1: Confusion matrices for one-class novelty detection using KNFST, KDE and SVDD using 2.5% of the training data. The red field represents the novel class.

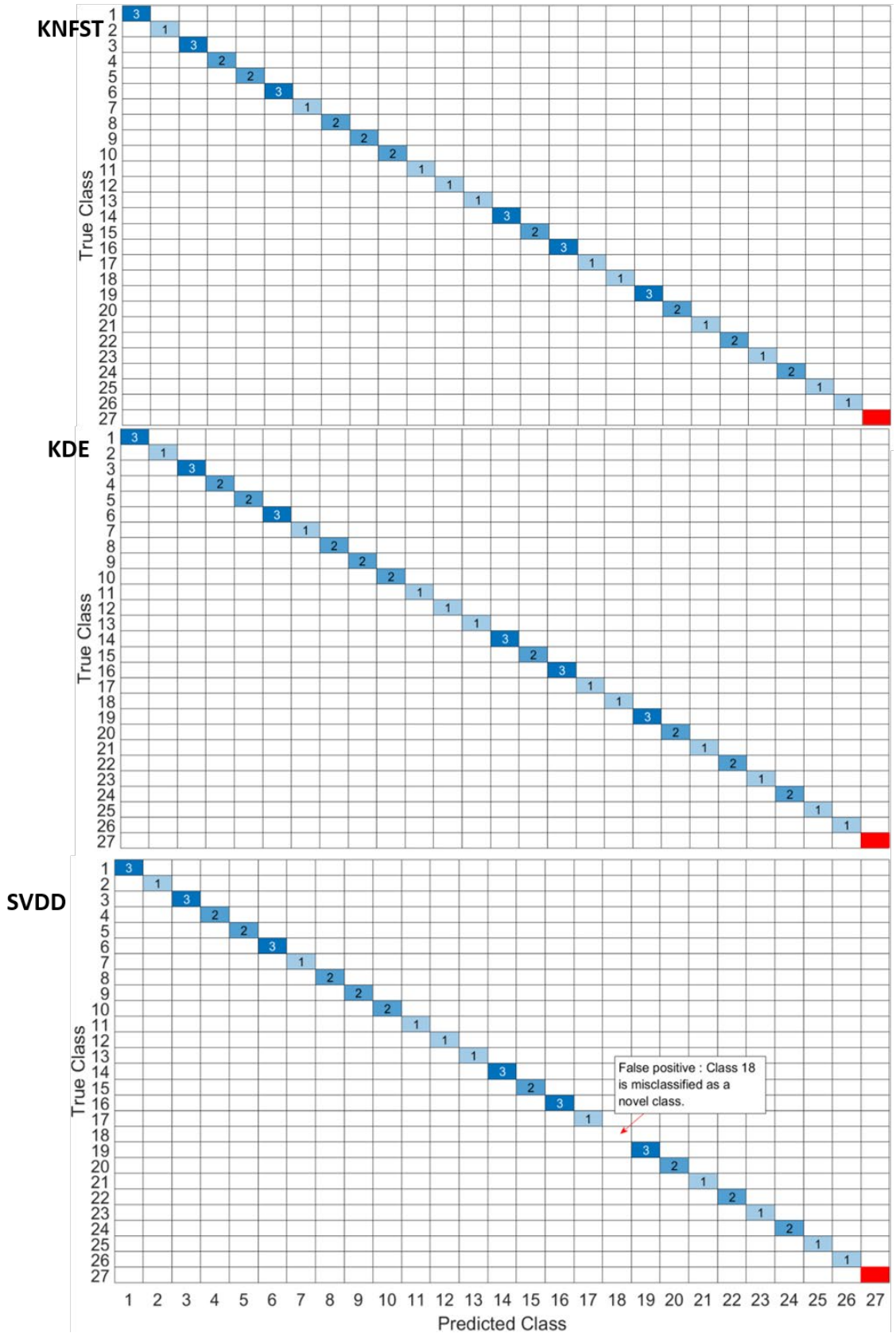


Figure Supp. 2: Confusion matrices for one-class novelty detection using KNFST, KDE and SVDD using 10% of the training data. The red field represents the novel class.

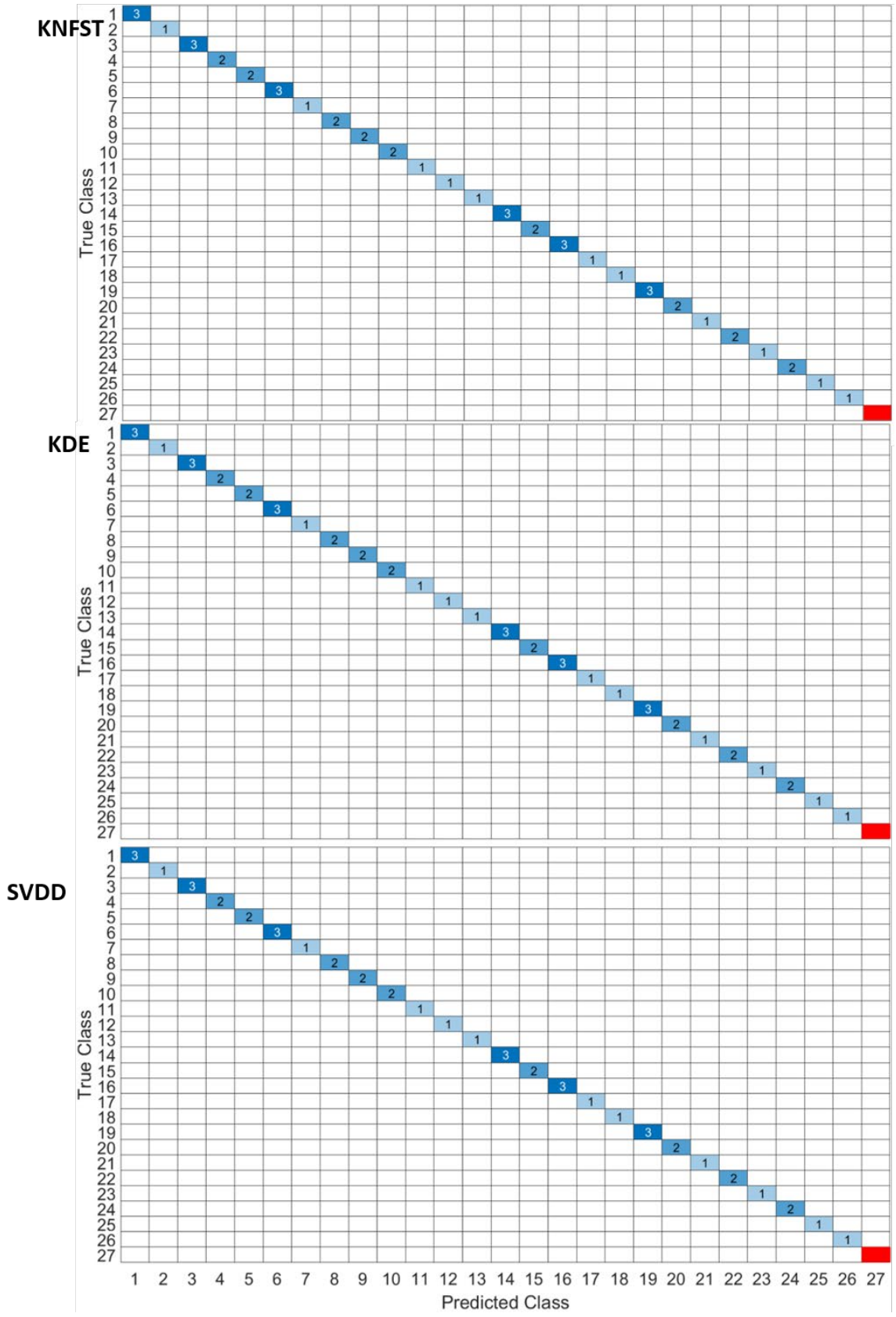


Figure Supp. 3: Confusion matrices for one-class novelty detection using KNFST, KDE and SVDD using 100% of the training data. The red field represents the novel class.

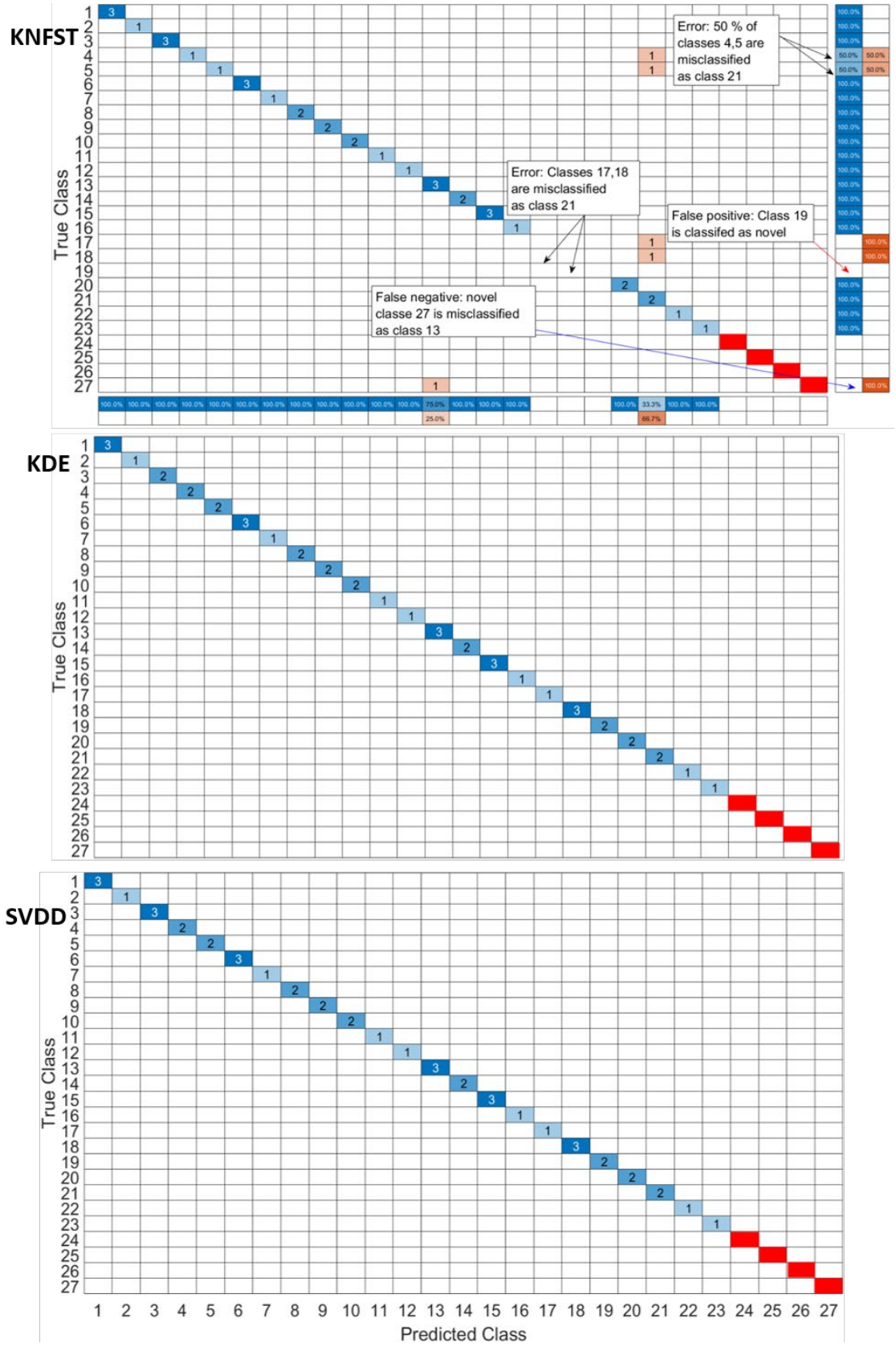


Figure Supp. 4: Confusion matrices for multi-class novelty detection using KNFST, KDE and SVDD using 2.5% of the training data. The red fields represent the novel classes.

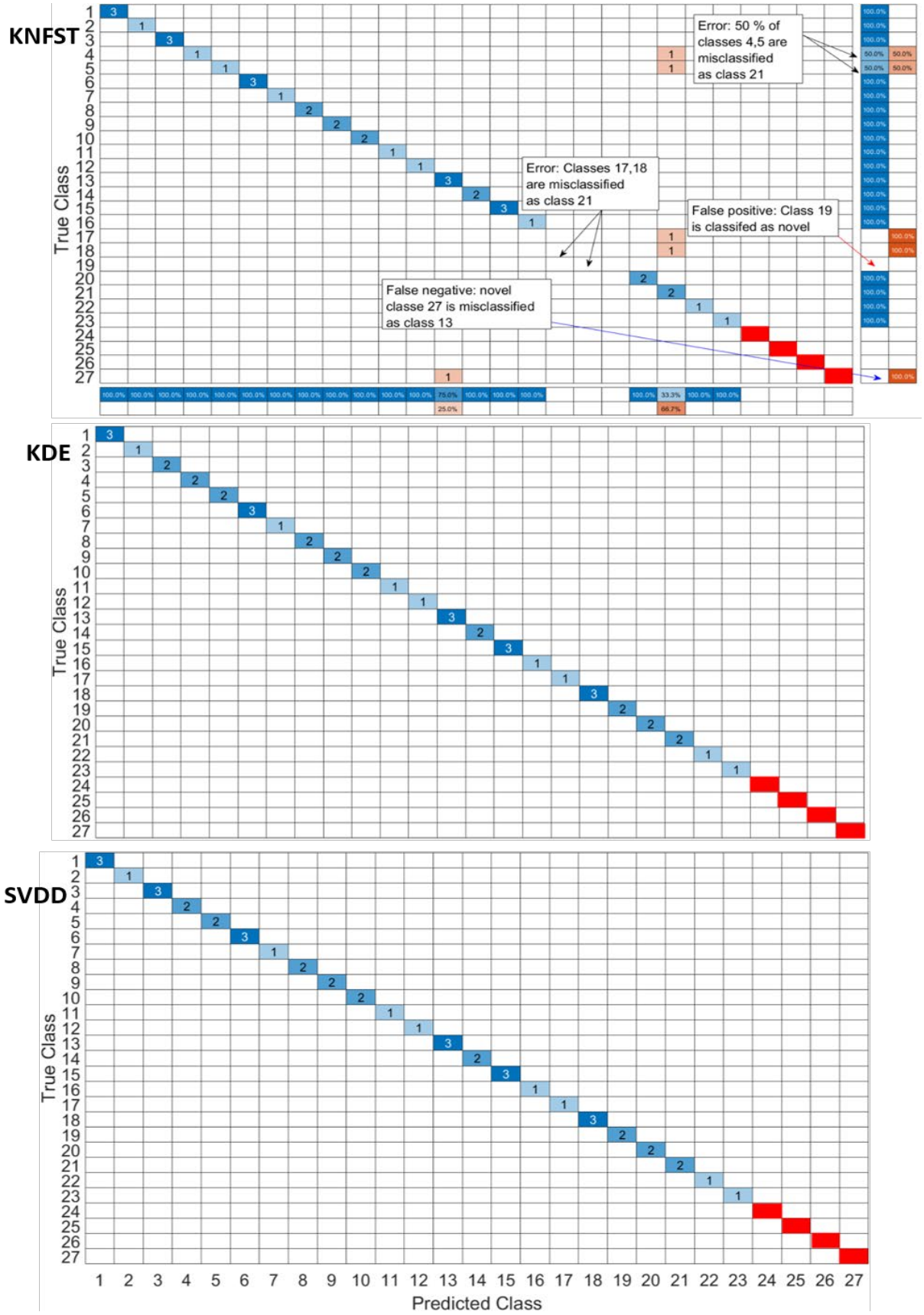


Figure Supp. 5: Confusion matrices for multi-class novelty detection using KNFST, KDE and SVDD using 10% of the training data. The red fields represent the novel classes.

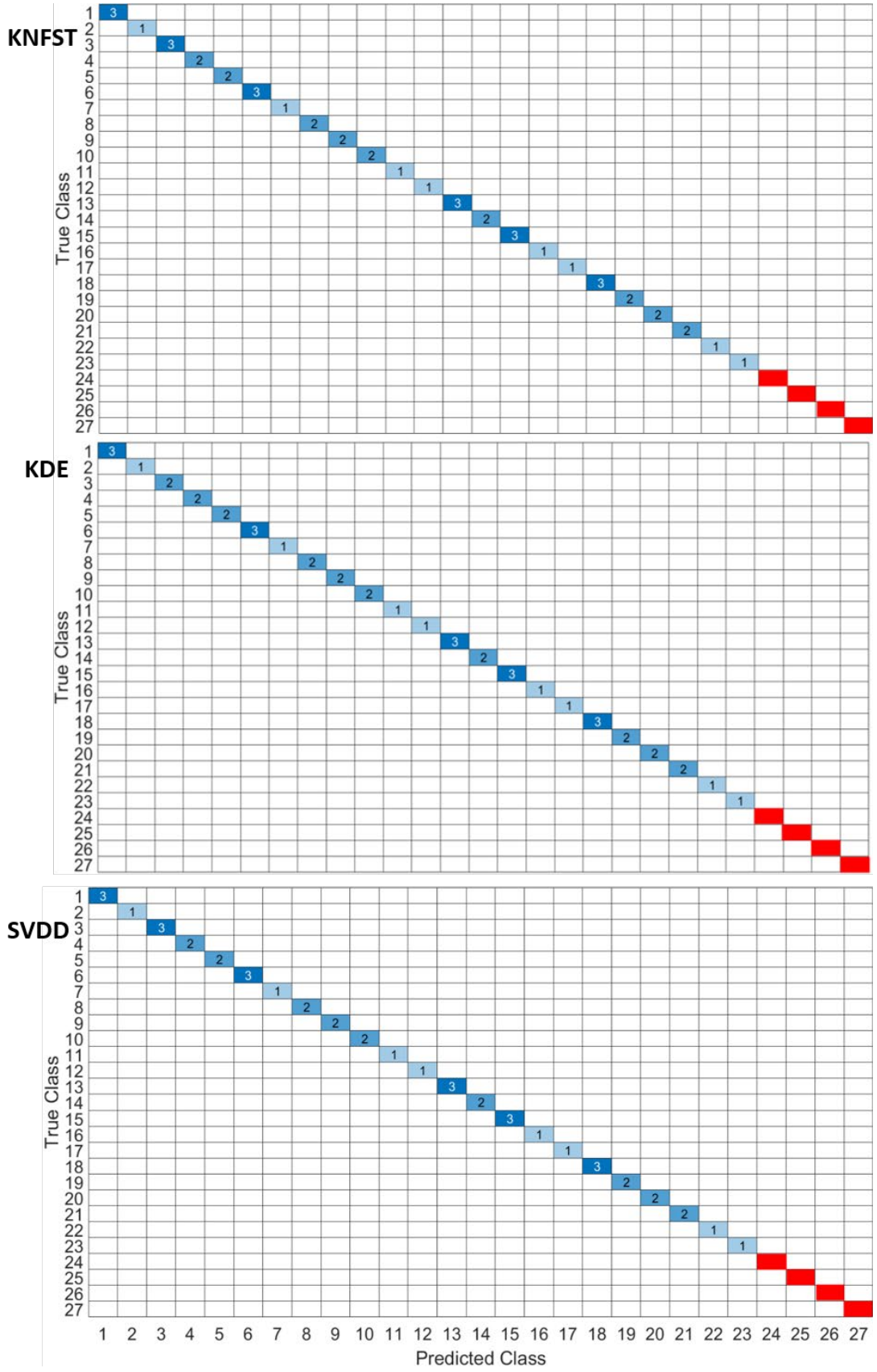


Figure Supp. 6: Confusion matrices for multi-class novelty detection using KNFST, KDE and SVDD using 100% of the training data.The red fields represent the novel classes.

The ROC curves and AUC values for metabolites proline, serine and leucine are shown in Figure Supp. 7 for the metabolites proline, serine, and leucine.

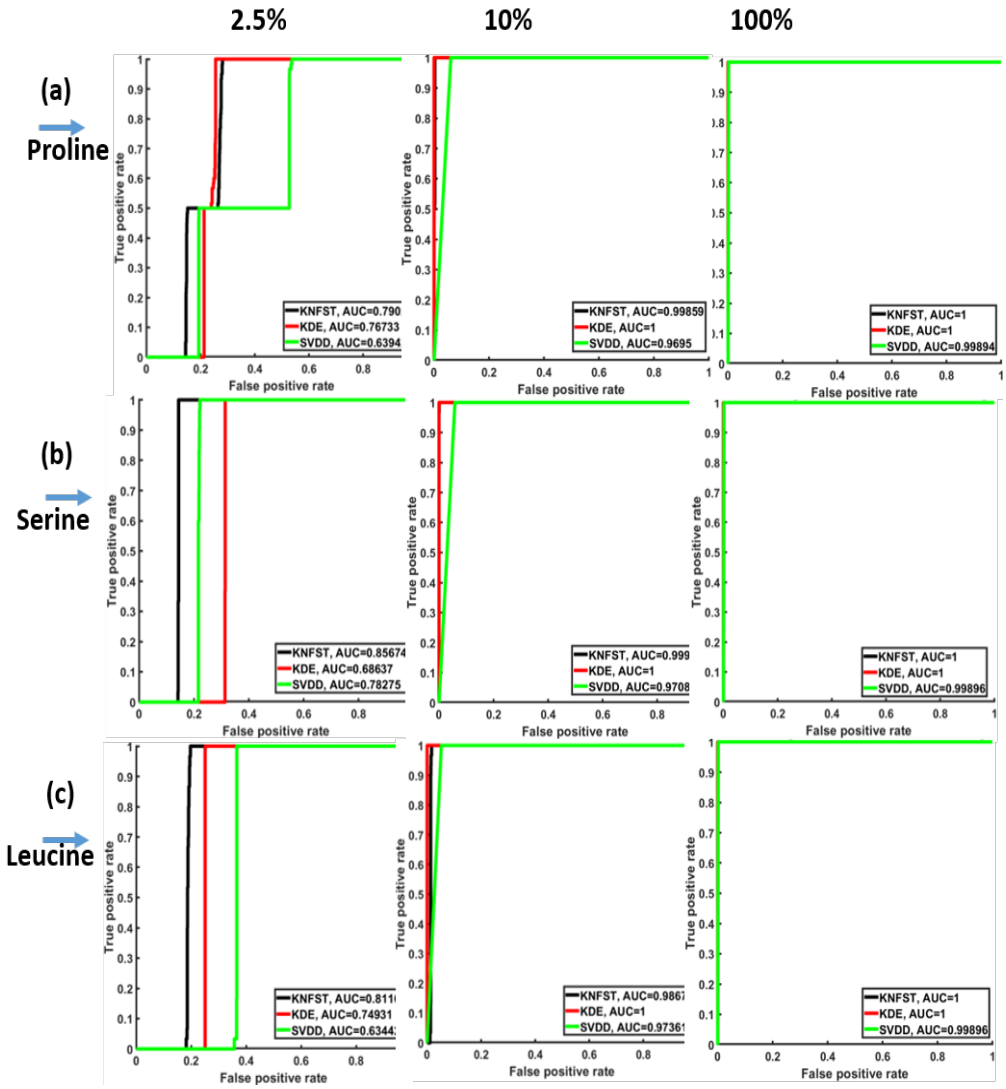


Figure Supp. 7: ROC curves and AUC values showing the accuracy of the novelty threshold for different sizes of training data for the metabolites proline, serine, and leucine. For each metabolite from left to right, the ROC curve using (a) 2.5%, (b) 10% and (c) 100% of the total size of the training dataset is shown for KNFST, (b) KDE and (c) SVDD as indicated in the subfigure's legends.

B. AT-derived hMSCs Sample preparation

i) Cultivation of AT-derived hMSCs

MSCs were maintained in basal MSCs culture media composed of alpha MEM medium with Earle's Salts (Gibco) supplemented with 5% human platelet lysate (hPL), at a concentration of 3 I.U Heparin-Sodium 5000 I.U/mL, 1% penicillin streptomycin, and 2 mM L-glutamine [294]. The cells were cultured in an adherent plate at a seeding density of 4000 cells/cm2, and subculture was performed every time the cells reached a confluence of 80% until reaching cell division in passage number 4 (P4). The passage

number indicates the number of times that cells have been collected and re-cultured into new cell culture flasks [295].

ii) Adipogenic and osteogenic differentiation of AT-derived hMSCs

AT-MSCs were induced to differentiate into adipocytes or osteocytes using StemPro Adipogenesis, Osteogenesis Differentiation Kit (Gibco), respectively as described by the manufacturer. In brief, MSCs at P4 were cultivated in MSCs basal culture media (BCM) at a seeding density of 4000 cells/cm2. When cells reached 70% confluence basal culture media was aspirated and cells were washed twice with PBS, before the addition of complete adipogenic (ADM) or osteogenic differentiation media (ODM). Cells were maintained in standard culture conditions (37 °C, 5% CO2) in humidified incubator for 14 days, while refeeding the cells every 3-4 days with complete fresh media. Through the differentiation duration, morphological changes in MSCs were monitored using inverted microscopy. To confirm the differentiation of MSCs into adipocytes and osteocytes at the end of the differentiation duration, the generated monolayer of adipogenic or osteogenic induced MSCs went through a staining procedure using oil red O for adipocytes, or Alizarin red staining for osteocytes [296] . Oil red staining illustrates the internal neutral lipids generated in adipocytes [297, 298], whereas alizarin red staining illustrate mineral deposits ,like calcium, generated by osteocytes [299]. BCM is supposed to maintain the stemness of MSCs without triggering their differentiation, this was confirmed by the lack of coloration in AT-Derived MSCs after 4 days of cultivation as seen on Figure Supp. 8a. Figure Supp. 8b shows AT-derived hMSCs that were cultured in basal cell culture media for 14 days, this media is supposed to maintain only their growth and stemness without triggering their differentiation. However, prolonged culture duration triggers the formation of lipid droplets (yellow to orange droplets). These cells were stained with both alizarin red and oil red stains, and the following was obtained: negative alizarin red staining, faded staining of oil red shown as yellow to orange droplets. It can be depicted on Figure Supp. 8c that MSCs cultivated in ADM for 14 days showed a clear alteration in their morphology due to the formation of large oil droplet in their cytoplasm as presented by the intense Oil red. Figure Supp. 8d shows osteogenic differentiation. MSCs cultivated in ODM exhibited an intense deposition of minerals, calcium, represented by the intense alizarin red staining.

iii) Intracellular metabolites extraction

At the end of the different periods, intracellular metabolites from the adipogenic and osteogenic differentiated AT-derived hMSCs plus their control group at 4 and 14 days of cultivation were extracted using methanol extraction method [300]. Briefly, differentiation media were aspirated, and the cultured cells were washed three times with phosphate-buffered saline (PBS). Immediately after washing, absolute methanol stored at −20 °C and water ice were added to the cells in a ratio of 2 parts:0.8 parts MeOH:H_2O to quench metabolism. Culture plates were stored at −80 °C for 10 min, then, the cells were scraped off the cell culture plate, and the obtained cells/methanol mixture were centrifuged at a speed of 14,000 rpm for 10 min. To obtain the intracellular metabolite

in powder form, the samples were lyophilized, and the obtained powder from each sample was stored at −80 °C until further use [296].

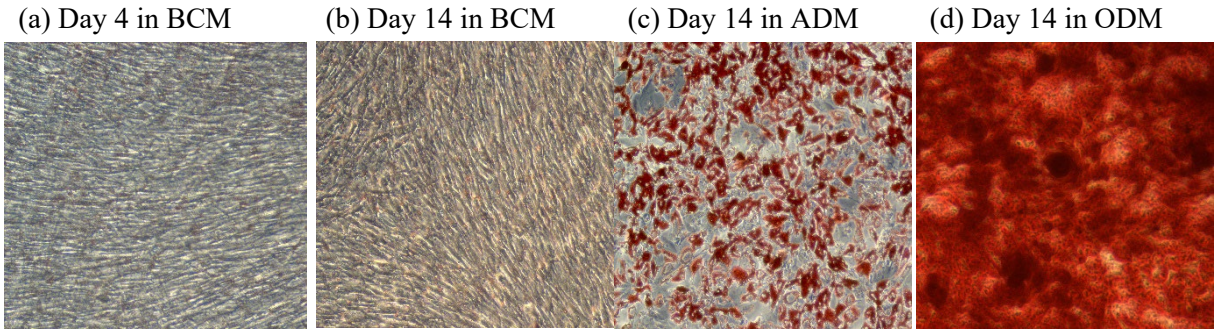


Figure Supp. 8: Light microscopy images of AT-derived hMSCs. (a) AT-derived hMSCs after 4 days, and (b) 14 days of cultivation in basal culture media (BCM). (c) Oil red staining illustrating adipogenic differentiation of AT-derived hMSCs after 14 days of cultivation in adipogenic differentiation media (ADM) [297]. (d) Alizarin red staining illustrating osteogenic differentiation of AT-derived hMSCs after 14 days of cultivation in osteogenic differentiation media (ODM) [299].

List of Figures

Figure 1.1: A noisy 2D NMR spectrum. Especially for samples with low intensity, NMR signal is contaminated by noise which appears as random fluctuating streaks in 2D NM resulting in reduced spectrum quality.....4

Figure 3.1: One-Dimensional NMR spectroscopy pulse-acquisition and Fourier transform. 19

Figure 3.2: Schematic diagram of ^{1}H chemical shift ranges for organic compounds 20

Figure 3.3 : ^{1}H NMR spectrum at 600.13 MHz of a HeLa cell extract showing metabolite annotated on the spectrum.....20

Figure 3.4: Two-Dimensional NMR spectroscopy pulse-acquisition. 21

Figure 3.5: 2D ^{1}H-^{1}H TOCSY contour plot of a urine sample.....23

Figure 4.1: A summary of the number of publications from years 2012 to 2021 on NMR in metabolomics. 26

Figure 5.1 : (a) The ^{1}H-^{1}H TOCSY spectrum of a breast cancer tissue sample.....34

Figure 5.2: Representative high resolution 1H NMR spectra of intracellular metabolite extracts obtained from AT-derived hMSC samples collected at day 14 of differentiation into adipocytes and osteocytes, and their control samples represented in AT-derived hMSC collected at day 4 and 14 of cultivation in BCM. 38

Figure 5.3: The feature space of the 27 metabolites deduced from the TOCSY spectrum of a breast cancer tissue. The magnifications are selected enlargements of peaks that overlap in (F1, F2) dimensions.....42

Figure 5.4: Feature space of the cross peaks of the metabolites contained in the samples (a) Ct d4, (b) Ct d14, (c) AT d14 and (d) OS d14. 42

Figure 6.1: Geometrical visualization of NFST. 48

Figure 6.2: Class membership in KNFST 50

Figure 6.3: The training phase in semi-supervised KNFST algorithm.....52

Figure 6.4: The learning phase in Semi-supervised KNFST algorithm.....53

Figure 6.5: The accuracy and mislabeling versus different sizes of initial training data.54

Figure 6.6: The accuracy and mislabeling versus different sizes of initial training data dataset for small initial amounts of labeled training data (≤9% of the entire dataset). 55

Figure 6.7: The metabolite assignment based on (a) the experimental 2D TOCSY NMR spectrum of breast cancer tissue after considering (b) the results of the KNFST classifier. 57

Figure 7.1: Schematic illustration of the ND procedure in metabolic profiling in a biological sample based on 2D TOCSY NMR spectra.....64

Figure 7.2: ND procedure by excluding one- and multi-metabolites from the pre-assigned 27 metabolites of the breast cancer tissue cell. 67

Figure 7.3: The Mnew, Fnew and Err values of breast cancer-tissue sample for the classifiers **(**a) KNFST, (b) KDE and (c) SVDD by applying one-class novelty detection. 69

Figure 7.4: ROC curves and AUC values showing the accuracy of the novelty threshold for different sizes of training data for the metabolite tyrosine. 70

Figure 7.5: Novelty scores and threshold values of KNFST, KDE and SVDD classifiers using different training dataset sizes in the one-class novelty detection.....70

Figure 7.6: Mnew, Fnew and Err values of breast cancer tissue sample for the classifiers (a) KNFST, (b) KDE and (c) SVDD by applying multi-class novelty detection. ..71

Figure 7.7: Novelty scores and threshold values of KNFST, KDE and SVDD classifiers for different training data size for multi-class novelty detection.....73

Figure 7.8: Total time from training to classification for (a) one-class and (b) multi-class novelty detection. 73

Figure 8.1: Schematic diagram of the experimental setting to observe the metabolic evolution of AT-derived hMSCs. 77

Figure 8.2: Confusion matrices of the output of classifiers KNFST and KDE for the spectrum after 14 days cultivation.....78

Figure 8.3: Confusion matrices of the output of classifiers KNFST and KDE for the spectrum of 14 days adiobocytes differentiation.....79

Figure 8.4: Confusion matrices of the output of classifiers KNFST and KDE for the spectrum of 14 days osteocytes differentiation.....80

Figure 8.5: Novelty scores and threshold values of KNFST and KDE classifiers for Ct d14 (a, b), AT d14(c, d) and OS d14(e, f). 81

Figure Supp. 1: Confusion matrices for one-class novelty detection using KNFST, KDE and SVDD using 2.5% of the training data.....87

Figure Supp. 2: Confusion matrices for one-class novelty detection using KNFST, KDE and SVDD using 10% of the training data.....88

Figure Supp. 3: Confusion matrices for one-class novelty detection using KNFST, KDE and SVDD using 100% of the training data. 89

Figure Supp. 4: Confusion matrices for multi-class novelty detection using KNFST, KDE and SVDD using 2.5% of the training data.....90

Figure Supp. 5: Confusion matrices for multi-class novelty detection using KNFST, KDE and SVDD using 10% of the training data.....91

Figure Supp. 6: Confusion matrices for multi-class novelty detection using KNFST, KDE and SVDD using 100% of the training data. 92

Figure Supp. 7: ROC curves and AUC values showing the accuracy of the novelty threshold for different sizes of training data for the metabolites proline, serine, and leucine. 93

Figure Supp. 8: Light microscopy images of AT-derived hMSCs. 95

List of Tables

Table 5.1: Breast cancer-tissue sample metabolites. 35

Table 5.2: Intracellular metabolites detected in AT-derived hMSCs at to control group at 4 days cultivation (Ct d4), 14 days of cultivation (Ct d14), 14 days of differentiation into adiobocytes (AT d14) and osteocytes (OS d14) and the standard frequencies from online libraries. 39

Table 5.3: A subset of the training dataset showing the output of the data augmentation procedure for tyrosine.....41

Table 6.1: Standard and experimental (F2, F1) Hz cross-peak entries of ^{1}H-^{1}H TOCSY of the metabolites appeared in the studied real breast cancer tissue. 58

Table 8.1: A summary of the performance of KDE and KNFST classifiers for Ct d14, AT d14 and OS d14.....81

List of Acronyms

AANN	Auto-Associative Networks
AT d14	AT-derived hMSCs after 14 days of adipocytes differentiation
AT- derived hMSCs	Adipose tissue-derived human MSCs
ATP	Adenosine Triphosphate
AUC	Area Under Cover
BMRB	Biological Magnetic Resonance Data Bank
CNN	Convolutional Neural Networks
COSY	Correlation Spectroscopy
CPMG	Carr-Purcell-Meiboom-Gill;
Ct d14	AT-derived hMSCs after 14 days of cultivation
Ct d4	AT-derived hMSCs after 4 days of cultivation
FID	Free Induction Decay
FT	Fourier transform
HMBC	Heteronuclear Multiple-Bond Correlation Spectroscopy
HMDB	Human Metabolome Database
HR MAS NMR	High-resolution magic angle spinning NMR
HSQC	Heteronuclear Single Quantum Coherence
J-RES	J-Resolved Spectroscopy
KDE	Kernel Density Estimation
KNFST	Kernel Null Foley–Sammon Transform
kNN	k- Neighbors Neural Network
ML	Machine learning
MLP	Multi-Layer Perceptron
MSCs	Mesenchymal stem cells
ND	Novelty Detection
NMR	Nuclear Magnetic Resonance
NN	Neural Networks
NOESY	Nuclear Overhauser Effect Spectroscopy
OS d14	AT-derived hMSCs after 14 days of osteocytes differentiation
PC	Polynomial Classifier
PCA	Principal Component Analysis
RBF	Radial Basis Function
RF pulse	Radio-Frequency Pulse
ROC	Receiver Operating Characteristic

ROESY	Rotating-Frame Nuclear Overhauser Effect Spectroscopy
SOM	Self-Organizing Networks
SSL	Semi-Supervised Learning
SVDD	Support Vector Data Description
SVM	Support Vector Machines
TMS	Tetramethylsilane
TOCSY	Total Correlation Spectroscopy
TSP	3-(trimethylsilyl) propionic acid sodium salt

Bibliography

1. Lindon, J.C., et al., *Metabonomics: metabolic processes studied by NMR spectroscopy of biofluids.* Concepts in Magnetic Resonance: An Educational Journal, 2000. **12**(5): p. 289-320.
2. Lin, L., et al., *Biomarkers of coordinate metabolic reprogramming and the construction of a co-expression network in colorectal cancer.* Ann Transl Med, 2022. **10**(20): p. 1115.
3. Gowda, G.A., et al., *Metabolomics-based methods for early disease diagnostics.* Expert Rev Mol Diagn, 2008. **8**(5): p. 617-33.
4. Mamas, M., et al., *The role of metabolites and metabolomics in clinically applicable biomarkers of disease.* Arch Toxicol, 2011. **85**(1): p. 5-17.
5. Lu, W., et al., *Metabolite Measurement: Pitfalls to Avoid and Practices to Follow.* Annu Rev Biochem, 2017. **86**: p. 277-304.
6. Davis, V.W., et al., *Metabolomics and surgical oncology: Potential role for small molecule biomarkers.* J Surg Oncol, 2011. **103**(5): p. 451-9.
7. Lee, M.Y. and T. Hu, *Computational Methods for the Discovery of Metabolic Markers of Complex Traits.* Metabolites, 2019. **9**(4): p. 66.
8. Manach, C., et al., *The complex links between dietary phytochemicals and human health deciphered by metabolomics.* Molecular nutrition & food research, 2009. **53**(10): p. 1303-1315.
9. Billoir, E., V. Navratil, and B.J. Blaise, *Sample size calculation in metabolic phenotyping studies.* Brief Bioinform, 2015. **16**(5): p. 813-9.
10. Zhang, A., et al., *Cell metabolomics.* OMICS, 2013. **17**(10): p. 495-501.
11. Emwas, A.H., et al., *NMR Spectroscopy for Metabolomics Research.* Metabolites, 2019. **9**(7).
12. Bingol, K., et al., *Unified and isomer-specific NMR metabolomics database for the accurate analysis of (13)C-(1)H HSQC spectra.* ACS Chem Biol, 2015. **10**(2): p. 452-9.
13. Hao, J., et al., *Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN.* Nat Protoc, 2014. **9**(6): p. 1416-27.
14. Gomez, J., et al., *Dolphin: a tool for automatic targeted metabolite profiling using 1D and 2D (1)H-NMR data.* Anal Bioanal Chem, 2014. **406**(30): p. 7967-76.
15. Zheng, C., et al., *Identification and quantification of metabolites in (1)H NMR spectra by Bayesian model selection.* Bioinformatics, 2011. **27**(12): p. 1637-44.
16. Peng, W.K., T.T. Ng, and T.P. Loh, *Machine learning assistive rapid, label-free molecular phenotyping of blood with two-dimensional NMR correlational spectroscopy.* Commun Biol, 2020. **3**(1): p. 535.
17. Peng, W.K., *Clustering Nuclear Magnetic Resonance: Machine learning assistive rapid two-dimensional relaxometry mapping.* Engineering Reports, 2021. **3**(10): p. e12383.
18. Wang, Y., X. Yu, and H. Zhao, *Biosystems design by directed evolution.* AIChE Journal, 2020. **66**(3): p. e16716.
19. de Matas, M., et al., *Strategic framework for education and training in Quality by Design (QbD) and process analytical technology (PAT).* Eur J Pharm Sci, 2016. **90**: p. 2-7.
20. Peng, B., H. Li, and X.X. Peng, *Functional metabolomics: from biomarker discovery to metabolome reprogramming.* Protein Cell, 2015. **6**(9): p. 628-37.

21. Schneider, G., *Automating drug discovery.* Nat Rev Drug Discov, 2018. **17**(2): p. 97-113.
22. Volk, M.J., et al., *Biosystems Design by Machine Learning.* ACS Synth Biol, 2020. **9**(7): p. 1514-1533.
23. Sommer, C. and D.W. Gerlich, *Machine learning in cell biology - teaching computers to recognize phenotypes.* J Cell Sci, 2013. **126**(Pt 24): p. 5529-39.
24. Mura, C., E.J. Draizen, and P.E. Bourne, *Structural biology meets data science: does anything change?* Curr Opin Struct Biol, 2018. **52**: p. 95-102.
25. Corsaro, C., et al., *NMR in Metabolomics: From Conventional Statistics to Machine Learning and Neural Network Approaches.* Applied Sciences, 2022. **12**(6).
26. Wu, D.S., *1D and 2D NMR Experiment Methods.* Emory University, 2011.
27. Harris, R.K., et al., *NMR nomenclature. Nuclear spin properties and conventions for chemical shifts(IUPAC Recommendations 2001).* Pure and Applied Chemistry, 2001. **73**(11): p. 1795-1818.
28. Mo, H., et al., *A simple method for NMR t(1) noise suppression.* J Magn Reson, 2017. **276**: p. 43-50.
29. Sengupta, A. and A.M. Weljie, *NMR Spectroscopy-Based Metabolic Profiling of Biospecimens.* Curr Protoc Protein Sci, 2019. **98**(1): p. e98.
30. Mitchell, T.M., *Machine learning. 1997.* Burr Ridge, IL: McGraw Hill, 1997. **45**(37): p. 870-877.
31. Hwang, T., *Computational Power and the Social Impact of Artificial Intelligence.* SSRN Electronic Journal, 2018.
32. Sukumar, S.R., *Machine Learning in the Big Data Era: Are We There Yet?* 2014.
33. Ching, T., et al., *Opportunities and obstacles for deep learning in biology and medicine.* J R Soc Interface, 2018. **15**(141): p. 20170387.
34. Bishop, C.M., *Pattern Recognition and Machine Learning* 2006, Berlin, Heidelberg: Springer-Verlag.
35. Murphy, K.P., *Machine Learning: A Probabilistic Perspective*. 2012: MIT Press.
36. Mohammed, M., M.B. Khan, and E.B.M. Bashier, *Machine Learning*. 2016, Boca Raton: CRC Press. 226.
37. Chapelle, O., B. Schölkopf, and A. Zien, *Semi-Supervised Learning*. 2010: The MIT Press.
38. Pimentel, M.A.F., et al., *A review of novelty detection.* Signal Processing, 2014. **99**: p. 215-249.
39. Roberts, S.J., *Novelty Detection using Extreme Value Statistics*, in *Vision, Image and Signal Processing*, 1999, IEE Proceedings. p. 124-129
40. Barber, D., *Bayesian Reasoning and Machine Learning*. 2012, Cambridge: Cambridge University Press.
41. Zhu, X. and A.B. Goldberg, *Introduction to Semi-Supervised Learning.* Synthesis Lectures on Artificial Intelligence and Machine Learning, 2009. **3**(1): p. 1-130.
42. Subramanya, A. and P.P. Talukdar. *Graph-Based Semi-Supervised Learning*. in *Graph-Based Semi-Supervised Learning*. 2014.
43. van Engelen, J.E. and H.H. Hoos, *A survey on semi-supervised learning.* Machine Learning, 2019. **109**(2): p. 373-440.
44. Graepel, T., R. Herbrich, and K. Obermayer, *Bayesian transduction.* Advances in Neural Information Processing Systems, 1999. **12**.
45. Triguero, I., S. García, and F. Herrera, *Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study.* Knowledge and Information Systems, 2013. **42**(2): p. 245-284.

46. Sawant, S.S. and M. Prabukumar, *A review on graph-based semi-supervised learning methods for hyperspectral image classification.* The Egyptian Journal of Remote Sensing and Space Science, 2020. **23**(2): p. 243-248.
47. Mills, M.T. and N.G. Bourbakis, *Graph-based methods for natural language processing and understanding—a survey and analysis.* IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2013. **44**(1): p. 59-71.
48. Blum, A. and T. Mitchell, *Combining labeled and unlabeled data with co-training*, in *Proceedings of the eleventh annual conference on Computational learning theory*, 1998, Association for Computing Machinery: Madison, Wisconsin, USA. p. 92-100
49. Kiritchenko, S. and S. Matwin. *Email classification with co-training*. in *CASCON*. 2001.
50. Xu, Q., et al., *Semi-supervised protein subcellular localization.* BMC Bioinformatics, 2009. **10 Suppl 1**(Suppl 1): p. S47.
51. Qiao, S., et al. *Deep Co-Training for Semi-Supervised Image Recognition*. in *Proceedings of the european conference on computer vision (eccv)*. 2018. Munich, Germany.
52. Romaszewski, M., P. Głomb, and M. Cholewa, *Semi-supervised hyperspectral classification from a small number of training samples using a co-training approach.* ISPRS Journal of Photogrammetry and Remote Sensing, 2016. **121**: p. 60-76.
53. Vapnik, V. and V. Vapnik, *Statistical learning theory Wiley.* New York, 1998. **1**: p. 624.
54. Bennett, K.P. and A. Demiriz, *Semi-supervised support vector machines*, in *Proceedings of the 11th International Conference on Neural Information Processing Systems*, 1998, MIT Press: Denver, CO. p. 368–374
55. Ahmed, F., M.F. Iqbal, and A. Rafiq, *On cone optimization approaches for semi-supervised support vector machines(S3VM).* 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 2019: p. 1-6.
56. Chapelle, O., V. Sindhwani, and S.S. Keerthi, *Optimization techniques for semi-supervised support vector machines.* Journal of Machine Learning Research, 2008. **9**: p. 203-233.
57. Maulik, U., A. Mukhopadhyay, and D. Chakraborty, *Gene-expression-based cancer subtypes prediction through feature selection and transductive SVM.* IEEE transactions on biomedical engineering, 2012. **60**(4): p. 1111-1117.
58. Joachims, T. *Transductive inference for text classification using support vector machines*. in *Sixteenth International Conference on Machine Learning*. 1999. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
59. Zemmal, N., et al. *Automated classification of mammographic abnormalities using transductive semi supervised learning algorithm*. in *Proceedings of the Mediterranean Conference on Information & Communication Technologies*. 2016. Springer.
60. Li, X., et al., *An object-based river extraction method via optimized transductive support vector machine for multi-spectral remote-sensing images.* IEEE Access, 2019. **7**: p. 46165-46175.
61. Goldberg, A.B., *New Directions in Semi-supervised Learning*, in *Computer Sciences*.PhD thesis, 2010, University of Wisconsin-Madison: Madison, Wisconsin, USA. p. 207
62. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society. Series B (Methodological), 1977. **39**(1): p. 1-38.
63. Zhang, W., X. Tang, and T. Yoshida, *Tesc: An approach to text classification using semi-supervised clustering.* Knowledge-Based Systems, 2015. **75**: p. 152-160.

64. Li, R., F.C. Pereira, and M.E. Ben-Akiva, *Competing risk mixture model and text analysis for sequential incident duration prediction.* Transportation Research Part C: Emerging Technologies, 2015. **54**: p. 74-85.
65. Nigam, K., et al., *Text Classification from Labeled and Unlabeled Documents using EM.* Machine Learning, 2000. **39**(2/3): p. 103-134.
66. Balafar, M., *Gaussian mixture model based segmentation methods for brain MRI images.* Artificial Intelligence Review, 2014. **41**(3): p. 429-439.
67. Wilson, R. *MGMM: multiresolution Gaussian mixture models for computer vision*. in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*. 2000.
68. Al-Behadili, H., A. Grumpe, and C. Wöhler. *Non-linear Distance-based Semi-supervised Multi-class Gesture Recognition*. in *VISIGRAPP (3: VISAPP)*. 2016.
69. Hillebrand, M., et al. *Semi-supervised Training Set Adaption to Unknown Countries for Traffic Sign Classifiers*. in *Partially Supervised Learning: First IAPR TC3 Workshop*. 2012. Ulm, Germany Springer.
70. Hillebrand, M., et al., *Traffic Sign Classifier Adaption by Semi-supervised Co-training*. Vol. 7477. 2012. 193-200.
71. Han, W., et al., *Semi-Supervised Active Learning for Sound Classification in Hybrid Learning Environments.* PLoS One, 2016. **11**(9): p. e0162075.
72. Schürmann, J., *Pattern classification: a unified view of statistical and neural approaches*. 1996: John Wiley & Sons, Inc.
73. Hillebrand, M., et al. *Self-learning with confidence bands*. in *Proc. 20th Workshop Computational Intelligence*. 2010. Citeseer.
74. Kendall, W.S., J.-M. Marin, and C.P. Robert, *Confidence bands for Brownian motion and applications to Monte Carlo simulation.* Statistics and Computing, 2007. **17**(1): p. 1-10.
75. Bluhmki, T., et al., *A wild bootstrap approach for the Aalen-Johansen estimator.* Biometrics, 2018. **74**(3): p. 977-985.
76. Cui, T., et al. *Analytically tractable sample-specific confidence measures for semi-supervised learning*. in *Proc. Workshop Computational Intelligence*. 2011.
77. Martos, A., L. Krüger, and C. Wöhler. *Towards Real Time Camera Self Calibration: Significance and Active Selection*. in *Proc. of the 4th Int. Symp. on 3D Data Processing, Visualization and Transmission (3DPVT)*. 2010.
78. Kardaun, O.J., *Classical methods of statistics: with applications in fusion-oriented plasma physics*. 2005: Springer Science & Business Media.
79. Moya, M.M. and D.R. Hush, *Network constraints and multi-objective optimization for one-class classification.* Neural networks, 1996. **9**(3): p. 463-474.
80. Khan, S.S. and M.G. Madden, *One-class classification: taxonomy of study and review of techniques.* The Knowledge Engineering Review, 2014. **29**(3): p. 345-374.
81. Markou, M. and S. Singh, *Novelty detection: a review—part 1: statistical approaches.* Signal Processing, 2003. **83**(12): p. 2481-2497.
82. Drews-Jr, P., et al., *Novelty detection and segmentation based on Gaussian mixture models: A case study in 3D robotic laser mapping.* Robotics and Autonomous Systems, 2013. **61**(12): p. 1696-1709.
83. Zhang, Y., et al., *Detection of emerging faults on industrial gas turbines using extended Gaussian mixture models.* International Journal of Rotating Machinery, 2017. **2017**.
84. Clifton, D.A., P.R. Bannister, and L. Tarassenko, *A framework for novelty detection in jet engine vibration data*, in *Key Engineering Materials*.Conference Paper, 2007, Trans Tech Publications Ltd. p. 305-310
85. Clifton, D.A., *Novelty Detection with Extreme Value Theory in Jet Engine Vibration Data*.PhD Thesis, 2009, St. Cross Colleg, University of Oxford

86. Clifton, D.A. and L. Tarassenko, *Novelty detection in jet engine vibration spectra.* International Journal of Condition Monitoring, 2015. **5**(2): p. 2-7.
87. Ntalampiras, S., I. Potamitis, and N. Fakotakis, *Probabilistic novelty detection for acoustic surveillance under real-world conditions.* IEEE Transactions on Multimedia, 2011. **13**(4): p. 713-719.
88. Lee, H. and S.J. Roberts. *On-line novelty detection using the Kalman filter and extreme value theory*. in *2008 19th International Conference on Pattern Recognition*. 2008.
89. Kapoor, A., et al., *Gaussian Processes for Object Categorization.* International Journal of Computer Vision, 2009. **88**(2): p. 169-188.
90. Tarassenko, L., A. Hann, and D. Young, *Integrated monitoring and analysis for early warning of patient deterioration.* Br J Anaesth, 2006. **97**(1): p. 64-8.
91. González, F.A. and D. Dasgupta, *Anomaly detection using real-valued negative selection.* Genetic Programming and Evolvable Machines, 2003. **4**(4): p. 383-403.
92. Gomez, J., F. Gonzalez, and D. Dasgupta. *An immuno-fuzzy approach to anomaly detection*. in *The 12th IEEE International Conference on Fuzzy Systems, 2003. FUZZ '03.* 2003.
93. Esponda, F., S. Forrest, and P. Helman, *A formal framework for positive and negative detection schemes.* IEEE Trans Syst Man Cybern B Cybern, 2004. **34**(1): p. 357-73.
94. Parzen, E., *On Estimation of a Probability Density Function and Mode.* The Annals of Mathematical Statistics, 1962. **33**(3): p. 1065-1076.
95. Hautamäki, V. and I. Karkkainen, *Outlier detection using k-nearest neighbour graph*. Vol. 3. 2004. 430-433 Vol.3.
96. Otey, M.E., A. Ghoting, and S. Parthasarathy, *Fast Distributed Outlier Detection in Mixed-Attribute Data Sets.* Data Mining and Knowledge Discovery, 2006. **12**(2-3): p. 203-228.
97. Clifton, D.A., P.R. Bannister, and L. Tarassenko. *Learning Shape for Jet Engine Novelty Detection*. 2006. Berlin, Heidelberg: Springer Berlin Heidelberg.
98. Schölkopf, B., et al., *Support vector method for novelty detection.* Advances in Neural Information Processing Systems 12, 2000. **12**: p. 582-588.
99. Tax, D.M.J. and R.P.W. Duin, *Support Vector Data Description.* Machine Learning, 2004. **54**(1): p. 45-66.
100. Shin, S.Y. and H.-j. Kim, *Extended Autoencoder for Novelty Detection with Reconstruction along Projection Pathway.* Applied Sciences, 2020. **10**(13): p. 4497.
101. Markou, M. and S. Singh, *Novelty detection: a review—part 2:: neural network based approaches.* Signal processing, 2003. **83**(12): p. 2499-2521.
102. Ghorbani, M.A., et al., *A comparative study of artificial neural network (MLP, RBF) and support vector machine models for river flow prediction.* Environmental Earth Sciences, 2016. **75**(6): p. 476.
103. Chandola, V., A. Banerjee, and V. Kumar, *Anomaly detection: A survey.* ACM computing surveys (CSUR), 2009. **41**(3): p. 1-58.
104. Schölkopf, B., A. Smola, and K.-R. Müller, *Nonlinear component analysis as a kernel eigenvalue problem.* Neural computation, 1998. **10**(5): p. 1299-1319.
105. Yu, N. and P. Jiao. *Handwritten digits recognition approach research based on distance & Kernel PCA*. in *2012 IEEE Fifth International Conference on Advanced Computational Intelligence (ICACI)*. 2012.
106. Akinnuwesi, B.A., B.O. Macaulay, and B.S. Aribisala, *Breast cancer risk assessment and early diagnosis using Principal Component Analysis and support vector machine techniques.* Informatics in Medicine Unlocked, 2020. **21**: p. 100459.
107. Zhang, B., et al., *Network Intrusion Detection Method Based on PCA and Bayes Algorithm.* Security and Communication Networks, 2018. **2018**: p. 1-11.

108. Fujimaki, R., T. Yairi, and K. Machida, *An approach to spacecraft anomaly detection problem using kernel feature space*, in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005, Association for Computing Machinery: Chicago, Illinois, USA. p. 401-410
109. Yu, G., et al. *Fault detection and diagnosis for spacecraft using principal component analysis and support vector machines*. in *2012 7th IEEE Conference on Industrial Electronics and Applications (ICIEA)*. 2012.
110. Fawcett, T., *ROC Graphs: Notes and Practical Considerations for Researchers.* Machine Learning, 2004. **31**: p. 1-38.
111. Clark, J., Z. Liu, and N. Japkowicz. *Adaptive Threshold for Outlier Detection on Data Streams*. 2018.
112. Van, Q.N., et al., *Comparison of 1D and 2D NMR spectroscopy for metabolic profiling.* J Proteome Res, 2008. **7**(2): p. 630-9.
113. Song, Z., et al., *Application of NMR metabolomics to search for human disease biomarkers in blood.* Clin Chem Lab Med, 2019. **57**(4): p. 417-441.
114. Günther, H., *NMR spectroscopy: basic principles, concepts and applications in chemistry*. 2013: John Wiley & Sons.
115. James, T.L., *Fundamentals of NMR.* Online Textbook: Department of Pharmaceutical Chemistry, University of California, San Francisco, 1998: p. 1-31.
116. Barron, A.R., *Physical methods in chemistry and nano science.* 2015.
117. Cujia, K.S., et al., *Tracking the precession of single nuclear spins by weak measurements.* Nature, 2019. **571**(7764): p. 230-233.
118. Louis-Joseph, A. and P. Lesot, *Designing and building a low-cost portable FT-NMR spectrometer in 2019: A modern challenge.* Comptes Rendus Chimie, 2019. **22**(9-10): p. 695-711.
119. Qu, X., et al., *Reconstruction of self-sparse 2D NMR spectra from undersampled data in the indirect dimension.* Sensors (Basel), 2011. **11**(9): p. 8888-909.
120. Hoffman, R.E., *Standardization of chemical shifts of TMS and solvent signals in NMR solvents.* Magn Reson Chem, 2006. **44**(6): p. 606-16.
121. Lenz, E.M., et al., *Cyclosporin A-induced changes in endogenous metabolites in rat urine: a metabonomic investigation using high field 1H NMR spectroscopy, HPLC-TOF/MS and chemometrics.* J Pharm Biomed Anal, 2004. **35**(3): p. 599-608.
122. Gartland, K.P., et al., *Pattern recognition analysis of high resolution 1H NMR spectra of urine. A nonlinear mapping approach to the classification of toxicological data.* NMR Biomed, 1990. **3**(4): p. 166-72.
123. Constantinou, M.A., et al., *1H NMR-based metabonomics for the diagnosis of inborn errors of metabolism in urine.* Analytica Chimica Acta, 2005. **542**(2): p. 169-177.
124. Sargsyan, G. *NMR Chemical Shift Values Table*. Organic Chemistry. Nuclear Magnetic Resonance (NMR) Spectroscopy [cited 2022 29.08.2022]; Available from: https://www.chemistrysteps.com/nmr-chemical-shift-values-table/.
125. Mili, M., et al., *Fast and ergonomic extraction of adherent mammalian cells for NMR-based metabolomics studies.* Anal Bioanal Chem, 2020. **412**(22): p. 5453-5463.
126. Simpson, J.H., *Organic structure determination using 2-D NMR spectroscopy: a problem-based approach*. 2011: Academic Press.
127. Keeler, J., *Understanding NMR Spectroscopy*. 2 ed. 2013: Wiley.
128. Macomber, R.S., *A complete introduction to modern NMR spectroscopy*. Vol. 4263. 1998: Wiley New York.
129. Aue, W.P., J. Karhan, and R.R. Ernst, *Homonuclear broad band decoupling and two-dimensional J-resolved NMR spectroscopy.* The Journal of Chemical Physics, 1976. **64**(10): p. 4226-4227.

130. Frydman, L., A. Lupulescu, and T. Scherf, *Principles and features of single-scan two-dimensional NMR spectroscopy.* J Am Chem Soc, 2003. **125**(30): p. 9204-17.
131. Berger, S., *Gradient Selected Constant Time Cosy.* Spectroscopy Letters, 2000. **33**(1): p. 1-8.
132. Chen, K., D.I. Freedberg, and D.A. Keire, *NMR profiling of biomolecules at natural abundance using 2D 1H-15N and 1H-13C multiplicity-separated (MS) HSQC spectra.* J Magn Reson, 2015. **251**: p. 65-70.
133. Bodenhausen, G. and D.J. Ruben, *Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy.* Chemical Physics Letters, 1980. **69**(1): p. 185-189.
134. Kay, L., P. Keifer, and T. Saarinen, *Pure absorption gradient enhanced heteronuclear single quantum correlation spectroscopy with improved sensitivity.* Journal of the American Chemical Society, 2002. **114**(26): p. 10663-10665.
135. Oman, T., et al., *Identification of metabolites from 2D (1)H-(13)C HSQC NMR using peak correlation plots.* BMC Bioinformatics, 2014. **15**(1): p. 413.
136. Fardus-Reid, F., J. Warren, and A. Le Gresley, *Validating heteronuclear 2D quantitative NMR.* Analytical Methods, 2016. **8**(9): p. 2013-2019.
137. Paudel, L., et al., *Simultaneously enhancing spectral resolution and sensitivity in heteronuclear correlation NMR spectroscopy.* Angew Chem Int Ed Engl, 2013. **52**(44): p. 11616-9.
138. Mauve, C., et al., *Sensitive, highly resolved, and quantitative (1)H-(13)C NMR data in one go for tracking metabolites in vegetal extracts.* Chem Commun (Camb), 2016. **52**(36): p. 6142-5.
139. Brüschweiler, R., et al., *Combined use of hard and soft pulses for ω1 decoupling in two-dimensional NMR spectroscopy.* Journal of Magnetic Resonance (1969), 1988. **78**(1): p. 178-185.
140. Yuwen, T. and N.R. Skrynnikov, *CP-HISQC: a better version of HSQC experiment for intrinsically disordered proteins under physiological conditions.* J Biomol NMR, 2014. **58**(3): p. 175-92.
141. Kiraly, P., et al., *Real-time pure shift (1)(5)N HSQC of proteins: a real improvement in resolution and sensitivity.* J Biomol NMR, 2015. **62**(1): p. 43-52.
142. Lane, A.N., *Principles of NMR for Applications in Metabolomics*, in *The Handbook of Metabolomics*. 2012, Humana Press: Totowa, NJ. p. 127-197.
143. Lindon, J.C., J.K. Nicholson, and E. Holmes, *The Handbook of Metabonomics and Metabolomics*. 1 ed. 2007: Elsevier.
144. Chylla, R.A., et al., *Deconvolution of two-dimensional NMR spectra by fast maximum likelihood reconstruction: application to quantitative metabolomics.* Anal Chem, 2011. **83**(12): p. 4871-80.
145. Chen, D., et al., *Review and Prospect: Deep Learning in Nuclear Magnetic Resonance Spectroscopy.* Chemistry, 2020. **26**(46): p. 10391-10401.
146. Cherni, A., et al., *Challenges in the decomposition of 2D NMR spectra of mixtures of small molecules.* Faraday Discuss, 2019. **218**(0): p. 459-480.
147. Snyder, D.A., F. Zhang, and R. Bruschweiler, *Covariance NMR in higher dimensions: application to 4D NOESY spectroscopy of proteins.* J Biomol NMR, 2007. **39**(3): p. 165-75.
148. Bingol, K. and R. Bruschweiler, *Multidimensional approaches to NMR-based metabolomics.* Anal Chem, 2014. **86**(1): p. 47-57.
149. Bingol, K. and R. Bruschweiler, *Deconvolution of chemical mixtures with high complexity by NMR consensus trace clustering.* Anal Chem, 2011. **83**(19): p. 7412-7.

150. Puchades-Carrasco, L., et al., *Bioinformatics tools for the analysis of NMR metabolomics studies focused on the identification of clinically relevant biomarkers.* Brief Bioinform, 2016. **17**(3): p. 541-52.
151. Boiteau, R.M., et al., *Structure Elucidation of Unknown Metabolites in Metabolomics by Combined NMR and MS/MS Prediction.* Metabolites, 2018. **8**(1): p. 8.
152. Markley, J.L., et al., *The future of NMR-based metabolomics.* Curr Opin Biotechnol, 2017. **43**: p. 34-40.
153. Crook, A.A. and R. Powers, *Quantitative NMR-Based Biomedical Metabolomics: Current Status and Applications.* Molecules, 2020. **25**(21): p. 5128.
154. Mahrous, E.A. and M.A. Farag, *Two dimensional NMR spectroscopic approaches for exploring plant metabolome: A review.* J Adv Res, 2015. **6**(1): p. 3-15.
155. Van, Q., et al., *Comparison of 1D and 2D NMR Spectroscopy for Metabolic Profiling.* Journal of proteome research, 2008. **7**: p. 630-9.
156. Chambers, E., et al., *Systematic and comprehensive strategy for reducing matrix effects in LC/MS/MS analyses.* J Chromatogr B Analyt Technol Biomed Life Sci, 2007. **852**(1-2): p. 22-34.
157. Fan, T.W. and A.N. Lane, *Applications of NMR spectroscopy to systems biochemistry.* Prog Nucl Magn Reson Spectrosc, 2016. **92-93**: p. 18-53.
158. Hansen, A.L., et al., *2D NMR-Based Metabolomics with HSQC/TOCSY NOAH Supersequences.* Anal Chem, 2021. **93**(15): p. 6112-6119.
159. Garcia-Perez, I., et al., *Identifying unknown metabolites using NMR-based metabolic profiling techniques.* Nat Protoc, 2020. **15**(8): p. 2538-2567.
160. Larive, C.K., G.A. Barding, Jr., and M.M. Dinges, *NMR spectroscopy for metabolomics and metabolic profiling.* Anal Chem, 2015. **87**(1): p. 133-46.
161. Ludwig, C. and M.R. Viant, *Two-dimensional J-resolved NMR spectroscopy: review of a key methodology in the metabolomics toolbox.* Phytochem Anal, 2010. **21**(1): p. 22-32.
162. Lewis, I.A., et al., *NMR method for measuring carbon-13 isotopic enrichment of metabolites in complex solutions.* Anal Chem, 2010. **82**(11): p. 4558-63.
163. van der Laan, T., et al., *Fractionation platform for target identification using off-line directed two-dimensional chromatography, mass spectrometry and nuclear magnetic resonance.* Anal Chim Acta, 2021. **1142**: p. 28-37.
164. Thrippleton, M.J. and J. Keeler, *Elimination of zero-quantum interference in two-dimensional NMR spectra.* Angew Chem Int Ed Engl, 2003. **42**(33): p. 3938-41.
165. Beckonert, O., et al., *Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts.* Nat Protoc, 2007. **2**(11): p. 2692-703.
166. Dona, A.C., et al., *A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments.* Comput Struct Biotechnol J, 2016. **14**: p. 135-53.
167. Reif, B., et al., *Solid-state NMR spectroscopy.* Nat Rev Methods Primers, 2021. **1**(1): p. 2.
168. Guntert, P., *Automated structure determination from NMR spectra.* Eur Biophys J, 2009. **38**(2): p. 129-43.
169. Williamson, M.P. and C.J. Craven, *Automated protein structure calculation from NMR data.* J Biomol NMR, 2009. **43**(3): p. 131-43.
170. Sugiki, T., et al., *Current NMR Techniques for Structure-Based Drug Discovery.* Molecules, 2018. **23**(1): p. 148.
171. Xia, J., et al., *MetaboAnalyst 3.0--making metabolomics more meaningful.* Nucleic Acids Res, 2015. **43**(W1): p. W251-7.

172. Wishart, D.S., et al., *HMDB 3.0--The Human Metabolome Database in 2013.* Nucleic Acids Res, 2013. **41**(Database issue): p. D801-7.
173. Kanehisa, M., et al., *Data, information, knowledge and principle: back to metabolism in KEGG.* Nucleic Acids Res, 2014. **42**(Database issue): p. D199-205.
174. Jewison, T., et al., *SMPDB 2.0: big improvements to the Small Molecule Pathway Database.* Nucleic Acids Res, 2014. **42**(Database issue): p. D478-84.
175. Bingol, K., et al., *Comprehensive Metabolite Identification Strategy Using Multiple Two-Dimensional NMR Spectra of a Complex Mixture Implemented in the COLMARm Web Server.* Anal Chem, 2016. **88**(24): p. 12411-12418.
176. Ulrich, E.L., et al., *BioMagResBank.* Nucleic Acids Res, 2008. **36**(Database issue): p. D402-8.
177. Elyashberg, M., *Identification and structure elucidation by NMR spectroscopy.* TrAC Trends in Analytical Chemistry, 2015. **69**: p. 88-97.
178. Elyashberg, M., A. Williams, and K. Blinov, *Contemporary Computer-Assisted Approaches to Molecular Structure Elucidation*. 2011.
179. Sheen, D.A., et al., *Chemometric Outlier Classification of 2D-NMR Spectra to Enable Higher Order Structure Characterization of Protein Therapeutics.* Chemometr Intell Lab Syst, 2020. **199**: p. 103973.
180. Brinson, R.G., et al., *Enabling adoption of 2D-NMR for the higher order structure assessment of monoclonal antibody therapeutics.* MAbs, 2019. **11**(1): p. 94-105.
181. Cheng, Y., X. Gao, and F. Liang, *Bayesian peak picking for NMR spectra.* Genomics Proteomics Bioinformatics, 2014. **12**(1): p. 39-47.
182. Gonczarek, A., et al. *A Bayesian Framework for Chemical Shift Assignment*. in *Asian Conference on Intelligent Information and Database Systems*. 2017. Springer.
183. Heinecke, A., et al., *Bayesian Deconvolution and Quantification of Metabolites from J-Resolved NMR Spectroscopy.* Bayesian Analysis, 2020.
184. Astle, W., et al., *A Bayesian Model of NMR Spectra for the Deconvolution and Quantification of Metabolites in Complex Biological Mixtures.* Journal of the American Statistical Association, 2012. **107**(500): p. 1259-1271.
185. Klukowski, P., et al., *Computer vision-based automated peak picking applied to protein NMR spectra.* Bioinformatics, 2015. **31**(18): p. 2981-8.
186. Kwon, Y., et al., *Neural Message Passing for NMR Chemical Shift Prediction.* J Chem Inf Model, 2020. **60**(4): p. 2024-2030.
187. Mathworks. *MATLAB Deep Learning Toolbox*. 2019 [cited]; Available from: https://www.mathworks.com/products/deep-learning.html.
188. Lee, H.H. and H. Kim, *Intact metabolite spectrum mining by deep learning in proton magnetic resonance spectroscopy of the brain.* Magn Reson Med, 2019. **82**(1): p. 33-48.
189. Kobayashi, N., et al., *Noise peak filtering in multi-dimensional NMR spectra using convolutional neural networks.* Bioinformatics, 2018. **34**(24): p. 4300-4301.
190. Kobayashi, N., et al., *KUJIRA, a package of integrated modules for systematic and interactive analysis of NMR data directed to high-throughput NMR structure studies.* J Biomol NMR, 2007. **39**(1): p. 31-52.
191. Yu, D., et al., *An introduction to computational networks and the computational network toolkit*, 2014, Microsoft Technical Report MSR-TR-2014–112
192. Buchner, L. and P. Guntert, *Systematic evaluation of combined automated NOE assignment and structure calculation with CYANA.* J Biomol NMR, 2015. **62**(1): p. 81-95.

193. Reher, R., et al., *A Convolutional Neural Network-Based Approach for the Rapid Annotation of Molecularly Diverse Natural Products.* J Am Chem Soc, 2020. **142**(9): p. 4114-4120.
194. Zhang, C., et al. *Small Molecule Accurate Recognition Technology (SMART) to Enhance Natural Products Research*. Scientific reports, 2017. **7**, 14243 DOI: 10.1038/s41598-017-13923-x.
195. Chopra, S., R. Hadsell, and Y. LeCun. *Learning a similarity metric discriminatively, with application to face verification*. in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. 2005. San Diego, CA, USA: IEEE.
196. Xu, Z.F., et al., *Human umbilical mesenchymal stem cell and its adipogenic differentiation: Profiling by nuclear magnetic resonance spectroscopy.* World J Stem Cells, 2012. **4**(4): p. 21-7.
197. Bispo, D.S.C., et al., *NMR Metabolomics Assessment of Osteogenic Differentiation of Adipose-Tissue-Derived Mesenchymal Stem Cells.* J Proteome Res, 2022. **21**(3): p. 654-670.
198. Bispo, D.S.C., et al., *Endo- and Exometabolome Crosstalk in Mesenchymal Stem Cells Undergoing Osteogenic Differentiation.* Cells, 2022. **11**(8): p. 1257.
199. Jang, M.Y., et al., *Evaluation of metabolomic changes as a biomarker of chondrogenic differentiation in 3D-cultured human mesenchymal stem cells using proton (1H) nuclear magnetic resonance spectroscopy.* PLoS One, 2013. **8**(10): p. e78325.
200. Foley, D.H. and J.W. Sammon, *An Optimal Set of Discriminant Vectors.* IEEE Transactions on Computers, 1975. **C-24**(3): p. 281-289.
201. Guo, Y.-F., et al., *Rapid and brief communication: Null Foley-Sammon transform.* Pattern Recognition, 2006. **39**(11): p. 2248-2251.
202. Castiglione, F., et al., *NMR Metabolomics for Stem Cell type discrimination.* Sci Rep, 2017. **7**(1): p. 15808.
203. Coope, A., et al., *(1)H NMR Metabolite Monitoring during the Differentiation of Human Induced Pluripotent Stem Cells Provides New Insights into the Molecular Events That Regulate Embryonic Chondrogenesis.* Int J Mol Sci, 2022. **23**(16): p. 9266.
204. Gogiashvili, M., et al., *Impact of intratumoral heterogeneity of breast cancer tissue on quantitative metabolomics using high-resolution magic angle spinning (1) H NMR spectroscopy.* NMR Biomed, 2018. **31**(2): p. e3862.
205. Migdadi, L., et al., *Automated metabolic assignment: Semi-supervised learning in metabolic analysis employing two dimensional Nuclear Magnetic Resonance (NMR).* Comput Struct Biotechnol J, 2021. **19**: p. 5047-5058.
206. Gogiashvili, M., et al., *HR-MAS NMR Based Quantitative Metabolomics in Breast Cancer.* Metabolites, 2019. **9**(2).
207. Mavel, S., et al., *1H-13C NMR-based urine metabolic profiling in autism spectrum disorders.* Talanta, 2013. **114**: p. 95-102.
208. Rai, R.K. and N. Sinha, *Fast and accurate quantitative metabolic profiling of body fluids by nonlinear sampling of 1H-13C two-dimensional nuclear magnetic resonance spectroscopy.* Anal Chem, 2012. **84**(22): p. 10005-11.
209. Roberts, L.D., et al., *Targeted metabolomics.* Curr Protoc Mol Biol, 2012. **Chapter 30**: p. Unit 30 2 1-24.
210. Ai, Z., et al., *Widely Targeted Metabolomics Analysis to Reveal Transformation Mechanism of Cistanche Deserticola Active Compounds During Steaming and Drying Processes.* Front Nutr, 2021. **8**: p. 742511.
211. Society, A.C., *Breast cancer facts & figures 2019–2020.* Am Cancer Soc, 2019: p. 1-44.

212. Ferlay, J., et al., *Cancer statistics for the year 2020: An overview.* Int J Cancer, 2021. **149**(4): p. 778-789.
213. Harbeck, N., et al., *Breast cancer.* Nat Rev Dis Primers, 2019. **5**(1): p. 66.
214. Levental, K.R., et al., *omega-3 polyunsaturated fatty acids direct differentiation of the membrane phenotype in mesenchymal stem cells to potentiate osteogenesis.* Sci Adv, 2017. **3**(11): p. eaao1193.
215. Assis-Ribas, T., et al., *Extracellular matrix dynamics during mesenchymal stem cells differentiation.* Dev Biol, 2018. **437**(2): p. 63-74.
216. Shi, C., et al., *HRMAS 1H-NMR measured changes of the metabolite profile as mesenchymal stem cells differentiate to targeted fat cells in vitro: implications for non-invasive monitoring of stem cell differentiation in vivo.* J Tissue Eng Regen Med, 2008. **2**(8): p. 482-90.
217. Kern, S., et al., *Artificial neural networks for quantitative online NMR spectroscopy.* Anal Bioanal Chem, 2020. **412**(18): p. 4447-4459.
218. Paruzzo, F.M., et al., *Chemical shifts in molecular solids by machine learning.* Nat Commun, 2018. **9**(1): p. 4501.
219. Liu, J., et al., *Deep convolutional neural networks for Raman spectrum recognition: a unified solution.* Analyst, 2017. **142**(21): p. 4067-4074.
220. Mikołajczyk, A. and M. Grochowski. *Data augmentation for improving deep learning in image classification problem*. in *International Interdisciplinary PhD Workshop (IIPhDW)*. 2018. Swinoujście.
221. Liu, S., et al., *Multiresolution 3D-DenseNet for Chemical Shift Prediction in NMR Crystallography.* J Phys Chem Lett, 2019. **10**(16): p. 4558-4565.
222. Bjerrum, E., M. Glahder, and T. Skov, *Data Augmentation of Spectral Data for Convolutional Neural Network (CNN) Based Deep Chemometrics.* ArXiv, 2017. **abs/1710.01927**.
223. Tredwell, G.D., et al., *Modelling the acid/base (1)H NMR chemical shift limits of metabolites in human urine.* Metabolomics, 2016. **12**(10): p. 152.
224. Rosenberg, C., M. Hebert, and H. Schneiderman. *Semi-supervised self-training of object detection models*. in *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05)*. 2005. Breckenridge, CO, USA29-36.
225. Wöhler, C., A. Berezhnoy, and R. Evans, *Estimation of elemental abundances of the lunar regolith using clementine UVVIS+NIR data.* Planetary and Space Science, 2011. **59**(1): p. 92-110.
226. Milani, N.S., et al. *Partially Supervised Gesture Recognition*. in *Proc. Workshop Computational Intelligence*. 2012. Dortmund, Germany.
227. Jeon, J.H. and Y. Liu. *Semi-supervised learning for automatic prosodic event detection using co-training algorithm*. in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 2009. Suntec, Singapore: Association for Computational Linguistics.
228. Kim, M., *Semi-supervised learning of hidden conditional random fields for time-series classification.* Neurocomputing, 2013. **119**: p. 339-349.
229. Culotta, A. and A. McCallum. *Confidence estimation for information extraction*. in *Proceedings of HLT-NAACL 2004: Short Papers*. 2004.
230. Settles, B., *Active learning literature survey*.Technical Report 2009, University of Wisconsin-Madison Department of Computer Sciences. p.
231. Smola, A.J. and B. Schölkopf, *A Tutorial on Support Vector Regression.* 2004. **14**: p. 199-222.

232. Chang, C.-C. and C.-J. Lin, *Libsvm.* ACM Transactions on Intelligent Systems and Technology, 2011. **2**(3): p. 1-27.
233. Zhang, B.-f., J.-s. Su, and X. Xu. *A Class-Incremental Learning Method for Multi-Class Support Vector Machines in Text Classification*. in *2006 International Conference on Machine Learning and Cybernetics*. 2006.
234. Hastie, T. and R. Tibshirani, *Classification by pairwise coupling.* Annals of statistics, 1998. **26**(2): p. 451-471.
235. Wahba, G., *Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV.* Advances in Kernel Methods-Support Vector Learning, 1999. **6**: p. 69-87.
236. Lin, H.-T., C.-J. Lin, and R.C. Weng, *A note on Platt's probabilistic outputs for support vector machines.* Machine Learning, 2007. **68**(3): p. 267-276.
237. Platt, J., *Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods.* Advances in large margin classifiers, 1999. **10**(3): p. 61-74.
238. Zheng, W., L. Zhao, and C. Zou, *Foley-Sammon optimal discriminant vectors using kernel approach.* IEEE Trans Neural Netw, 2005. **16**(1): p. 1-9.
239. Duda, R.O., P.E. Hart, and D.G. Stork, *Pattern classification*. 2012: John Wiley & Sons.
240. Lin, Y., et al. *Kernel null foley-sammon transform*. in *2008 International Conference on Computer Science and Software Engineering*. 2008. IEEE.
241. Bodesheim, P., et al., *Kernel Null Space Methods for Novelty Detection.* Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2013.
242. Zhang, L., T. Xiang, and S. Gong. *Learning a discriminative null space for person re-identification*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
243. Schölkopf, B., A. Smola, and K.-R. Müller, *Kernel principal component analysis*, in *Artificial Neural Networks — ICANN'97*. 1997, Springer Berlin Heidelberg. p. 583-588.
244. Abdi, H., *The eigen-decomposition: Eigenvalues and eigenvectors.* Encyclopedia of measurement and statistics, 2007: p. 304-308.
245. Liu, W., et al., *Null space approach of fisher discriminant analysis for face recognition.* Biometric Authentication, Proceedings, 2004. **3087**: p. 32-44.
246. Guo, J., et al., *Smartphone-Based Patients' Activity Recognition by Using a Self-Learning Scheme for Medical Monitoring.* J Med Syst, 2016. **40**(6): p. 140.
247. Good, I., *What are degrees of freedom?* The American Statistician, 1973. **27**(5): p. 227-228.
248. Hall, P. and M.A. Martin, *A note on the accuracy of bootstrap percentile method confidence intervals for a quantile.* Statistics & probability letters, 1989. **8**(3): p. 197-200.
249. Pfeuffer, J., et al., *Toward an in vivo neurochemical profile: quantification of 18 metabolites in short-echo-time (1)H NMR spectra of the rat brain.* J Magn Reson, 1999. **141**(1): p. 104-20.
250. Govindaraju, V., K. Young, and A.A. Maudsley, *Proton NMR chemical shifts and coupling constants for brain metabolites.* NMR Biomed, 2000. **13**(3): p. 129-53.
251. Gogiashvili, M., *Quantitatives, nicht gezieltes metabolisches Profiling von Brustkrebsgewebe mittels HR-MAS NMR-Spektrometrie: analytische Aspekte und Zusammenhänge mit klinisch-pathologischen Parametern*. 2018: Westfälische Wilhelms-Universität Münster.

252. Migdadi, L., et al., *Novelty detection for metabolic dynamics established on breast cancer tissue using 2D NMR TOCSY spectra.* Comput Struct Biotechnol J, 2022. **20**: p. 2965-2977.
253. Wang, J., et al. *Evaluating features for person re-identification*. in *2016 IEEE International Conference on Signal and Image Processing (ICSIP)*. 2016. IEEE.
254. Luo, Y., et al., *Manifold learning for novelty detection and its application in gesture recognition.* Complex & Intelligent Systems, 2022: p. 1-12.
255. Shi, Y., et al., *Kernel null-space-based abnormal event detection using hybrid motion information.* Journal of Electronic Imaging, 2019. **28**(2): p. 021011.
256. Oza, P. and V.M. Patel. *Federated Learning-based Active Authentication on Mobile Devices*. in *2021 IEEE International Joint Conference on Biometrics (IJCB)*. 2021.
257. Tian, Y., et al., *A subspace learning-based feature fusion and open-set fault diagnosis approach for machinery components.* Advanced Engineering Informatics, 2018. **36**: p. 194-206.
258. Khan, S.S. and M.G. Madden. *A Survey of Recent Trends in One Class Classification*. in *Artificial Intelligence and Cognitive Science*. 2010. Berlin, Heidelberg: Springer Berlin Heidelberg.
259. Clifton, L.A., *Multi-channel novelty detection and classifier combination*.PhD thesis, 2007, The University of Manchester (United Kingdom): Manchester, United Kingdom
260. Chenge, D., et al. *Gesture Classification Algorithm Based on SVDD-EPF*. in *2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP)*. 2021.
261. Zhihong, Z., et al. *One-class classification for spontaneous facial expression analysis*. in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*. 2006.
262. Al-Behadili, H., et al., *Incremental Class Learning and Novel Class Detection of Gestures Using Ensemble*. 2015.
263. Yoshida, T. and T. Kitamura. *Semi-Hard Margin Support Vector Machines for Personal Authentication with an Aerial Signature Motion*. in *Artificial Neural Networks and Machine Learning – ICANN 2021*. 2021. Cham: Springer International Publishing.
264. Na, X.G., *New medical image classification approach based on hypersphere multi-class support vector data description.* Journal of Computer Applications, 2013. **33**(11): p. 3300.
265. Belghith, A., C. Collet, and J.P. Armspach. *Detection of Biomarker in Biopsies Based on Hr-Mas 2D HSQC Spectroscopy Indexation*. in *4th International Conference on Biomedical Engineering in Vietnam*. 2013. Berlin, Heidelberg: Springer Berlin Heidelberg.
266. Zhao, Y., S. Wang, and F. Xiao, *Pattern recognition-based chillers fault detection method using support vector data description (SVDD).* Applied Energy, 2013. **112**: p. 1041-1048.
267. Chen, M.-C., et al., *An efficient ICA-DW-SVDD fault detection and diagnosis method for non-Gaussian processes.* International Journal of Production Research, 2016. **54**(17): p. 5208-5218.
268. Qin, X., et al. *Scalable Kernel Density Estimation-based Local Outlier Detection over Large Data Streams*. in *22nd International Conference on Extending Database Technology (EDBT)*. 2019. Lisbon.
269. Wu, S., et al., *Automated fibroglandular tissue segmentation and volumetric density estimation in breast MRI using an atlas-aided fuzzy C-means method.* Med Phys, 2013. **40**(12): p. 122302.

270. Veluppal, A., et al., *Automated differentiation of Alzheimer's condition using Kernel Density Estimation based texture analysis of single slice brain MR images.* Current Directions in Biomedical Engineering, 2021. **7**(2): p. 747-750.
271. Yan, H., et al., *Volumetric magnetic resonance imaging classification for Alzheimer's disease based on kernel density estimation of local features.* Chin Med J (Engl), 2013. **126**(9): p. 1654-60.
272. Sadhukhan, D., et al., *Lateral ventricle texture analysis in alzheimer brain mr images using kernel density estimation.* Biomed Sci Instrum, 2021. **57**: p. 2.
273. Sarv Ahrabi, S., et al., *Exploiting probability density function of deep convolutional autoencoders' latent space for reliable COVID-19 detection on CT scans.* J Supercomput, 2022. **78**(9): p. 12024-12045.
274. Patel, A., et al., *Cross Attention Transformers for Multi-modal Unsupervised Whole-Body PET Anomaly Detection.* 2022.
275. Clifton, D.A., et al., *Automated novelty detection in industrial systems*, in *Advances of Computational Intelligence in Industrial Systems*. 2008, Springer. p. 269-296.
276. Bishop, C.M., *Neural networks for pattern recognition*
1995, Birmingham, UK: Clarendon Press. 504.
277. Masud, M., et al., *Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints.* IEEE Transactions on Knowledge and Data Engineering, 2011. **23**(6): p. 859-874.
278. Yang, L., et al., *Application of metabolomics in the diagnosis of breast cancer: a systematic review.* Journal of Cancer, 2020. **11**(9): p. 2540-2551.
279. Mandrekar, J.N., *Receiver operating characteristic curve in diagnostic test assessment.* J Thorac Oncol, 2010. **5**(9): p. 1315-6.
280. Gramacki, A., *Nonparametric kernel density estimation and its computational aspects*. Studies in Big Data. 2018: Springer Cham.
281. Peredriy, S., D. Kakde, and A. Chaudhuri. *Kernel bandwidth selection for SVDD: The sampling peak criterion method for large data*. in *2017 IEEE International Conference on Big Data (Big Data)*. 2017. Boston, USA.
282. Bishop, C.M., *Novelty detection and neural network validation.* IEE Proceedings-Vision, Image and Signal processing, 1994. **141**(4): p. 217-222.
283. Chaudhuri, A., et al. *Sampling Method for Fast Training of Support Vector Data Description*. in *2018 Annual Reliability and Maintainability Symposium (RAMS)*. 2018. IEEE.
284. Dufrenois, F. and J.-C. Noyer. *A null space based one class kernel Fisher discriminant*. in *2016 International Joint Conference on Neural Networks (IJCNN)*. 2016. IEEE.
285. Saeedi, P., R. Halabian, and A.A. Imani Fooladi, *A revealing review of mesenchymal stem cells therapy, clinical perspectives and Modification strategies.* Stem Cell Investig, 2019. **6**: p. 34.
286. Andrzejewska, A., B. Lukomska, and M. Janowski, *Concise Review: Mesenchymal Stem Cells: From Roots to Boost.* Stem Cells, 2019. **37**(7): p. 855-864.
287. Chu, D.T., et al., *Adipose Tissue Stem Cells for Therapy: An Update on the Progress of Isolation, Culture, Storage, and Clinical Application.* J Clin Med, 2019. **8**(7): p. 917.
288. Funes, J.M., et al., *Transformation of human mesenchymal stem cells increases their dependency on oxidative phosphorylation for energy production.* Proc Natl Acad Sci U S A, 2007. **104**(15): p. 6223-8.
289. Rocha, B., et al., *Metabolic labeling of human bone marrow mesenchymal stem cells for the quantitative analysis of their chondrogenic differentiation.* J Proteome Res, 2012. **11**(11): p. 5350-61.

290. Liu, Y. and T. Ma, *Metabolic regulation of mesenchymal stem cell in expansion and therapeutic application.* Biotechnol Prog, 2015. **31**(2): p. 468-81.
291. Salazar-Noratto, G.E., et al., *Understanding and leveraging cell metabolism to enhance mesenchymal stem cell transplantation survival in tissue engineering and regenerative medicine applications.* Stem Cells, 2020. **38**(1): p. 22-33.
292. Zhu, H., et al., *Inducible metabolic adaptation promotes mesenchymal stem cell therapy for ischemia: a hypoxia-induced and glycogen-based energy prestorage strategy.* Arterioscler Thromb Vasc Biol, 2014. **34**(4): p. 870-6.
293. Migdadi, L., et al., *Machine Learning in Automated Monitoring of Metabolic Changes Accompanying the Differentiation of Adipose-Tissue-Derived Human Mesenchymal Stem Cells Employing 1H-1H TOCSY NMR.* Metabolites, 2023. **13**(3): p. 352.
294. Abuarqoub, D., A. Awidi, and N. Abuharfeil, *Comparison of osteo/odontogenic differentiation of human adult dental pulp stem cells and stem cells from apical papilla in the presence of platelet lysate.* Arch Oral Biol, 2015. **60**(10): p. 1545-53.
295. England, P.H., *Passage numbers explained*, in *Culture Collections*, 2017, Public Health England
296. Ghorbani, A., S.A. Jalali, and M. Varedi, *Isolation of adipose tissue mesenchymal stem cells without tissue destruction: a non-enzymatic method.* Tissue Cell, 2014. **46**(1): p. 54-8.
297. Sathishkumar, S., P. Mohanashankar, and P. Boopalan, *Cell surface protein expression of stem cells from human adipose tissue at early passage with reference to mesenchymal stem cell phenotype.* Int J Med Med Sci, 2011. **3**(5): p. 129-134.
298. Zhang, A.X., et al., *Proteomic identification of differently expressed proteins responsible for osteoblast differentiation from human mesenchymal stem cells.* Mol Cell Biochem, 2007. **304**(1-2): p. 167-79.
299. Umrath, F., et al., *iPSC-Derived MSCs Versus Originating Jaw Periosteal Cells: Comparison of Resulting Phenotype and Stem Cell Potential.* Int J Mol Sci, 2020. **21**(2): p. 587.
300. Martineau, E., et al., *Strategy for choosing extraction procedures for NMR-based metabolomic analysis of mammalian cells.* Anal Bioanal Chem, 2011. **401**(7): p. 2133-42.