Special Problem Proposal

PanCan Spectrum: A Pancreatic Cancer Detection Support Tool Using Mass Spectrometry Data and Support Vector Machines

Emmanuel A. Briones

Department of Physical Sciences and Mathematics

University of the Philippines Manila

Adviser: Geoffrey Solano

June 2019

Abstract

Pancreatic cancer is one of the most fatal types of cancer due to its difficulty of being diagnosed in the early stages. Presently, multiple screening procedures for this disease are required to determine its presence. In this study, a pancreatic detection support tool implementing machine learning is to be created with support vector machines (SVM) algorithm and mass spectrometry data of pancreatic cancer patients and controls as training and testing datasets. The final output would aid researchers in detecting pancreatic cancer in patients (complementing current and common procedures), and in finding biomarkers of the disease.

Keywords: mass spectrometry, machine learning, pancreatic cancer detection, biomarkers, support vector machines

Contents

Abstract **ii**

List of Figures **v**

List of Tables **vi**

I. Introduction **1**

A. Background of the Study 1

B. Statement of the Problem 2

C. Objectives of the Study 3

D. Significance of the Project 4

E. Scope and Limitations 5

F. Assumptions 5

II. Review of Related Literature **7**

III. Theoretical Framework **12**

A. Definition of Terms 12

B. Pancreatic Cancer 12

C. Mass Spectrometry 14

D. Machine Learning 15

E. Supervised Learning 16

F. Classification in Machine Learning 16

G. MS Data Pre-processing 17

1. Noise Reduction 17

2. Data Normalization 18

3. Data Reduction 19

4. Peak Alignment 20

H. Support Vector Machines 21
I. Evaluation of Classifier 23
1. K-Fold Cross Validation 23
2. Model Evaluation Metrics 23

IV. Design and Implementation **25**
A. Datasets 25
B. System Architecture 25
1. Development Flow Diagram 25
2. Context Diagram 27
3. Use Case Diagrams 27
4. Data Flow Diagrams 29
5. Entity Relationship Diagram 31
C. Technical Architecture 31

V. Expected Output **33**

VI. Schedule of Activities **34**

VII. Bibliography **35**

List of Figures

1 Machine Learning Development Flow Diagram 26

2 PCDST Context Diagram 27

3 Use Case Diagram of Researcher / Oncologist 28

4 Use Case Diagram of Statistician 28

5 Top-Level Dataflow Diagram of PCDST 29

6 Dataflow Diagram of Process 1 30

7 Dataflow Diagram of Process 2 30

8 Dataflow Diagram of Processes 3-5 30

9 Entity Relationship Diagram of Chemical Compounds Database 31

List of Tables

I. Introduction

A. Background of the Study

Pancreatic cancer is a type of disease in which malignant cells form in the tissues of the pancreas. It is the 12th most common type of cancer with around 460,000 cases in 2018 [1] and is also one of the deadliest with an estimated survival rate of 9% [2]. There are several pancreatic cancer risk factors. Those that are lifestyle-related include smoking, obesity, being overweight, and workplace exposure to certain chemicals; while the other factors are family history, inherited gene syndromes, diabetes, and age. The risk of developing pancreatic cancer increases as people get older as studies show that almost all patients are older than 45 [3].

In pancreatic cancer diagnosis, there are multiple tests that are administered by oncologists in order to detect the presence of the disease and determine the extent of the condition of a particular patient [4]. Some examples of these tests are Magnetic Resonance Imaging (MRI), Computed Tomography (CT) Scan, Ultrasound, Cholangiopancreatography, blood tests, and Positron Emission Tomography (PET) Scan.

In the present, information technology and computer science play a major role in numerous fields including medicine, biochemistry, and toxicology. Many studies and researches have integrated techniques such as machine learning algorithms, statistical methods, and scientific processes in the identification and classification of diseases, including cancer. Among these techniques is mass spectrometry or mass spectroscopy.

Mass spectrometry (MS), or mass spectroscopy, is an analytical, instrumental technique used for the process of separating electrically charged species in a given biological sample using a mass spectrometer (which then produces a mass spectrum of mass/charge values with corresponding intensities) [5]. This particular technique has a lot of appli-

cations in the medical sciences. Some of the common uses of mass spectrometry are the diagnosis of various diseases and identification of the corresponding biomarkers. Furthermore, it has been greatly recognized due to its applications in drug safety evaluation and diagnosis of diseases [6].

B. Statement of the Problem

Pancreatic cancer detection can be difficult since the disease does not usually cause many specific symptoms in the early stages. Symptoms such as loss of appetite, indigestion, jaundice and abdominal pain can also be caused by more common illnesses such as pancreatitis, irritable bowel syndrome, and gallstones [7].

In the present, pancreatic cancer detection is based only on the clinicopathological attributes of a patient. Moreover, patients undergo several procedures such as ultrasound scans, CT scans, biopsies, MRI scans in order to detect the presence of pancreatic cancer. Although there exist many tests for diagnosing pancreatic cancer, no major professional group recommend routine screening for pancreatic cancer in patients who are at average risk because no screening test has been proven to lower the risk of death from pancreatic cancer [8]. This study aims to make the detection process easier and faster but at the same time reliable through the use of mass spectrometry data and machine learning. The results of the final system can also be used in the validation of other detection procedures mentioned beforehand.

Furthermore, the study seeks to answer the following questions:

- What are the common mass-per-charge values with high intensities found in the mass spectra of pancreatic cancer patients?
- Which chemical compounds (based on the mass/charge ratio values in a mass spectrum) are most commonly associated to pancreatic cancer?

- How are the mass-per-charge values associated to pancreatic cancer identified?

C. Objectives of the Study

This study aims to create a pancreatic cancer detection support tool which implements machine learning (SVM) using mass spectrometry data, in order to aid researchers in determining whether a patient has pancreatic cancer based on his or her mass spectrum.

- The researcher should be able to:
  - Input and file parse mass spectrometry data using the tool
  - Pre-process mass spectrometry data and perform the following:
    - Baseline Reduction
    - Smoothing
    - Data Normalization
    - Peak Reduction
    - Peak Alignment
  - View visualization of inputted mass spectrometry data
  - View corresponding numerical values of mass spectrum attributes
  - Identify the most abundant chemical compounds in a patient's sample based on the intensities of the mass/charge values of a spectrum
  - Perform pancreatic cancer detection using mass spectrometry data
  - Download report of detection result in text format
  - View user's manual
- The statistician should be able to:
  - Input and file parse mass spectrometry data using the tool

- Pre-process mass spectrometry data and perform the following:
  - Baseline Reduction
  - Smoothing
  - Data Normalization
  - Peak Reduction
  - Peak Alignment
- View visualization of inputted mass spectrometry data
- View corresponding numerical values of mass spectrum attributes
- Add inputted data as training and testing dataset to create a classification model
- View user's manual

D. Significance of the Project

The proposed system would enable oncologists to detect early stages of pancreatic cancer using mass spectrometry data acquired from biological samples extracted from patients. Moreover, biomarker identification for pancreatic cancer based on mass spectra mass-per-charge values with high intensities can be done.

The application of machine learning on mass spectrometry would help researchers in their discovery of new information regarding chemical compounds that are present in patients diagnosed with pancreatic cancer, extending their knowledge with regard to diagnostics of the disease hence giving vital contribution both to the scientific and medical community.

Furthermore, the system to be developed would serve as an aid for those working with mass spectrometry data, seeing that they could utilize it with regard to the establish-

ment of hypotheses regarding the correlation among numerous chemical compounds and pancreatic cancer, based on the system's outputted results. Ultimately, the system would lessen the work of oncologists with regard to running multiple screening tests and it would help them decide with regard to a patient's treatment.

E. Scope and Limitations

The limitations of the system are the following:

- The model would be trained based only on mass spectrometry data of pancreatic cancer-diagnosed patients, and of controls who do not have the disease.
- The output of the system does not indicate the type of pancreatic cancer based on the mass spectrometry data.
- Evaluation of the SVM model is based on accuracy, precision, recall, and f1-score after cross validation.
- The included mass spectrometry data pre-processing steps are to be used based on user's own selection.
- The tables in the chemical compounds database cannot be updated.

F. Assumptions

The assumptions in the study are the following:

- The datasets used for creating the machine learning model (training and testing) are acquired from the Global Natural Products Social Molecular Networking (GNPS) MassIVE Database and the PRoteomics IDEntifications (PRIDE) Archive, all of which are provided by researches and professionals working on studies related to mass spectrometry and other analytical techniques that involve natural products.

- The mass spectrometry datasets vary with regard to what type of mass spectrometer was used for acquisition.
- The data including the identification of chemical compounds are acquired from the Chemical Entities of Biological Interest (ChEBI) database, which is also contributed to by research professionals.
- The mass spectrometry data to be used for the system are in *mzml* format, which is the standard format for mass spectra (mass spectrometer output).
- The types of pancreatic cancer in the mass spectrometry dataset used for creating the machine learning model are not considered.
- The mass spectrometry data to be used are pre-processed using the tool before proceeding to the disease detection functionality.

II. Review of Related Literature

Some methodologies, researches, and concepts that relate to the proposed system would be discussed in this section. In brief, the following contains review of literature concerning mass spectrometry, machine learning, and classification algorithms.

Mass spectrometry is an analytical technique for the characterization of biological molecules and is increasingly used due to its targeted, nontargeted, and high throughput abilities. Machine learning has been applied to mass spectrometry data in the past from different biological disciplines, particularly for various cancers. The objectives of such investigations have been to identify biomarkers and to aid in detection, prognosis, and treatment of particular diseases [9].

In 2013, Swan et al. discussed the use of machine learning applied to proteomics data for identification of biomarkers. It was shown that there are several algorithms suitable for classification of samples and identification of biomarkers (i.e. Support Vector Machines, Decision Trees, Bayesian classifiers, Artificial Neural Networks, etc.), hence a study requiring the optimal method based on the objective and available resources involves the consideration of several important matters. These include the quantity of data to be used for training, the type of data may it be mass spectral peak data or identified proteins, and whether biomarker identification is needed besides unknown sample classification [9].

A study performed by Rajapakse et al. in 2005 shows that there are several basic steps involved before machine learning could be applied to mass spectrometry data with regard to cancer classification and biomarker discovery. First involves the collection and analysis of mass spectrometry data, then followed by baseline subtraction and noise reduction, which are necessary before peak extraction and alignment could be performed. Once the datasets have been prepared, partitioning into training and test sets is done. Afterwards,

the peak selection on training sets and the classification and validation is finally executed [10].

A typical data set used in a clinical application of mass spectrometry contains tens or hundreds of spectra; each spectrum contains thousands of intensity measurements representing an unknown number of protein peaks. Any attempt to make sense of this volume of data requires extensive low-level processing in order to identify the locations of peaks and to quantify their sizes accurately. Inadequate or incorrect preprocessing methods, however, can result in data sets that exhibit substantial biases and make it difficult to reach meaningful biological conclusions [11].

In 2006, Smith et al. introduced a tool called XCMS (various forms (X) Chromatography Mass Spectrometry) which incorporates the needed steps in MS data pre-processing mentioned. The three highlighted aspects of XCMS are its availability (being open-source), design, and flexibility. Moreover, it includes easily integrated tools for MS data pre-processing, analysis, and can be customized and tweaked for varying MS data analysis purposes or optimized for a specific application. Lastly, even if XCMS has been designed for LC/MS data, it could be modified to be able to use other data types [12].

One of mass spectrometry's strengths is its potential for biomarker discovery, which is one of the reasons why it has been used in medical and biochemical research for a long time [13]. Despite its potential, it is recognized that the acquisition of significant proteomic features from mass spectrometry data requires precise assessment. Consequently, a good feature selection method is needed.

A research done by Datta in 2013 involved the use of several algorithms including Random Forests (RF), Support Vector Machine (SVM), Partial Least Squares (PLS), Linear and Quadratic Discriminant Analysis (LDA and QDA, respectively). A performance comparison for the five algorithms was provided and the results show that PLS

has the highest overall accuracy and sensitivity, while SVM has the highest specificity or the ability to identify negative results (True Negative). Furthermore, it was stated that SVM has the advantage over other classifiers when it comes to flexibility, and has low classification error rates and robustness to several types of features [13].

Another study that involved the use and comparison of several machine learning techniques for feature selection was done by Ahmed et al. in 2013. The mainly proposed approach is the use of genetic programming (GP) for both feature selection and classification of mass spectrometry data. In this approach, two feature selection metrics are utilized: Information Gain (IG), which dictates the total amount of information gained with regard to a class when a particular feature existent or not; and Relief-F (REFS-F), which locates the two closest neighbors for a specific example: one from the same class (hit) and the other from a different class (miss); and then computes the value of the feature. Subsequently, the performance of the proposed method was analyzed and contrasted with IG and REFS-F on five mass spectrometry datasets with varying number of instances and features. Naïve-Bayes (NB), Support Vector Machines (SVM), and J48 Decision Trees (J48) are used to calculate the selected features' classification accuracy. It was concluded in the study that Genetic Programming as a feature selection method can pick out lesser number of features with better classification accuracy than IG and REFS-F using Naïve-Bayes, Support Vector Machines, and J48 DT. Moreover, GP surpasses the performance of NB and J48 as a classification method, and has minorly better performance than SVMs on the data sets used [14].

A study research done by Kim et al. in 2014 involved the use of various classification methods such as Random Forests, Bagging, Lasso, and Classification and Regression Tree (CART). Using mass spectrometry data of blood samples taken from patients with diabetes and pancreatic cancer, they were able to identify biomarkers linked to the diseases. Their methodology included the typical MS data pre-processing, use of various

classification algorithms, and cross validations [15].

In 2015, Nguyen et al. proposed a hybrid feature extraction method for cancer diagnosis using mass spectrometry data. After using Haar wavelets as features to convert mass spectrometry data into wavelet coefficients, genetic algorithm (GA) is applied to acquire feature sets from the most prominent wavelets. Using a variety of performance metrics such as accuracy, F-measure, and area under the curve (AUC), the Wavelet-GA methodology was shown to hold a significant advantage compared to other feature extraction methods such as Wilcoxon Test, t-test, principal component analysis (PCA), sequential search, etc. with regard to robustness [16]. Furthermore, a cross-validation was applied to three benchmark mass spectrometry datasets in the study, validating the conclusions that were formulated from the study. In conclusion, the Wavelet-GA can be implemented to create classification models that are to be used by researchers and medical experts for decision support systems in the field of oncology.

Another study that includes several feature selection techniques was done by Wong et al. in 2008. The researchers used Student t test, Wilcoxon rank sum test, and genetic algorithm to reduce the high dimensionality of pancreatic cancer MS data. With the features selected from each method, the performances of decision-tree based classifier ensembles were compared with that of a single decision-tree algorithm. The results show that classifier ensembles have better accuracies compared to single decision-tree [17].

Aside from the accuracy rate of feature selection and classification algorithms, another factor to be considered is the robustness to noise and outliers in a specific mass spectrometry dataset. In 2006, Zhang et al. developed a recursive support vector machine (R-SVM) algorithm to be used for the selection of biomarkers that are significant for classification of noisy data. The performance of the algorithm was compared to SVM Recursive Feature Elimination (SVM-RFE), focusing on both the algorithms' ability to mark truly informa-

tional biomarkers, and the robustness to the mass spectrometry dataset's outliers. After conducting the research, it was found out that R-SVM had 5%-20% better performance than SVM-RFE with regard to the features aforementioned. The R-SVM algorithm was subsequently implemented to two proteomics datasets: one regarding a breast cancer study and another from a research on rat liver cirrhosis. After the biomarkers have been found and validated, it was stated that the R-SVM algorithm is appropriate for evaluating both proteomics and microarray data which contains a relatively high amount of noise [18]. Furthermore, the proposed method performs better than SVM-RFE when it comes to gathering informative feature in the data and robustness to data noise.

Also emphasizing an algorithm's robustness to noise is a study by Pham et al. in 2011. The proposal involves another feature extraction method that considers the noise in a particular mass spectrometry data. The method used in the study integrates stationary wavelet transformation (SWT) and bivariate shrinkage estimator for denoising and feature extraction from MS data. Subsequently, two types of statistical feature testing, including Kolmogorov–Smirnov (KS) test and Mann–Whitney U (MW) test, are implemented on denoised wavelet coefficients in order to correctly pick the significant features that would then be made use of for identification of biomarkers. For method performance evaluation with regard to cancer classification, the researchers used a double-cross validation SVM classifier, accentuated to have high generalizability; and a Modest AdaBoost classifier which has a significantly better runtime. After application of both methods on Matrix-assisted laser desorption/ionization—time-of-flight (MALDI-TOF) datasets, the results of the research show that although the proposed method has better runtime than SVM, the latter still outperforms the former when it comes to sensitivity (True Positive) and specificity (True Negative) rates [19].

III. Theoretical Framework

A. Definition of Terms

1. **Mass Spectrum** - an intensity vs. mass-to-charge (m/z) plot of a chemical sample, acquired using a mass spectrometer [20].

2. **Ionization** – the process in which a molecule or an atom gains a positive or negative charge through losing or gaining electrons to form ions, usually happening with other chemical changes [21].

3. **Biomarker** – measurable substance from a patient which shows indications of a particular medical state, may it be a disease, infection, abnormal condition, etc. [22].

B. Pancreatic Cancer

Pancreatic cancer is a type of disease in which cancer cells form in the tissues of the pancreas [4]. It begins in the tissues of the pancreas — an organ the human abdomen that lies horizontally behind the lower part the stomach. The pancreas releases enzymes that aid digestion and hormones that help blood sugar levels [23].

Pancreatic cancer typically spreads rapidly to nearby organs. It is seldom detected in its early stages, but for people with pancreatic cysts or a family history of pancreatic cancer, some screening steps might help detect a problem early [23]. This type of disease is also considered to be a silent one since there are not many noticeable symptoms in its early stages. As the cancer develops, symptoms may include [24]:

- Yellow skin and eyes, darkening of the urine, itching, and clay-colored stool, which are signs of jaundice caused by a blockage of the bile ducts

- Pain in upper abdomen or upper back

- Burning feeling in stomach or other gastrointestinal discomforts
- Loss of appetite
- Nausea and vomiting
- Unexplained weight loss

Currently, there are no proven biomarkers that could make the early detection of pancreatic cancer more efficient. Nevertheless, researchers are striving to discover biomarkers that could indicate whether a person may have undiagnosed pancreatic cancer [25]. Many of the projects aimed at finding biomarkers to detect pancreatic cancer earlier focus on blood samples, comparing blood from patients who have the disease to healthy individuals.

There are multiple tests that are administered to patients in order to detect pancreatic cancer. These include [26]:

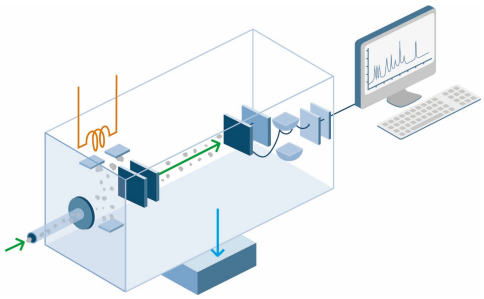
- Imaging tests such as Computerized Tomography (CT) scans, Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) scans
- Endoscopic Ultrasound (EUS)
- Removing a tissue sample for testing (biopsy)
- Blood tests

After a patient has been diagnosed with pancreatic cancer, he/she may undergo the following treatments [27]:

- Chemotherapy or Radiation Therapy
- Whipple procedure (pancreaticoduodenectomy)
- Distal pancreatectomy
- Total pancreatectomy

C. Mass Spectrometry

Mass spectrometry (MS), or mass spectroscopy, is an analytical type of instrumental technique used for the process of separating electrically charged species in a particular sample, using a mass spectrometer [28].



The mass spectrometry process can be sectioned into three main parts: the ionization, the analyzation, and the detection [5]. A particular sample is put into the ionization source of the mass spectrometer. Subsequently, once in the ionisation source, the molecules of the sample are ionized. This is done because ions are relatively easier to manipulate compared to neutral (neither positively or negatively charged) molecules. Afterwards, the ions are extracted into the analyzer part of the mass spectrometer in which the process of separation according to mass-to-charge (m/z) ratio is done. Lastly, the categorized ions are then detected and each signal (peak) is sent to a data system such that the m/z ratio data are inputted along with their corresponding relative abundance, all in a plot format called a mass spectrum.

A mass spectrum can be viewed as a histogram which provides data on the quantity of ions at varying values of mass-to-charge ratio (m/z) [28]. The detected ions indicate the presence of molecules or other chemical species developed over the process of ionization. Hence, mass spectrometry allows the identification of molecules based on the mass-to-charge ratio and its corresponding intensity, as well as the fragmentation patterns.

Mass spectrometry has a lot of applications in the fields of medicinal science. In medicine, some of the common uses of mass spectrometry are the diagnosis of many types of diseases and identification of the corresponding biomarkers. The utilization of mass spectrometry, together with multiple types of chromatography, is also being continuously developed due to its applications in drug safety evaluation and diagnosis of diseases [6]. Furthermore, the technique is used in a wide scope of applications, ranging from cancer diagnostics to forensic toxicology [29].

D. Machine Learning

Machine learning (ML) is a core subdiscipline of artificial intelligence (AI) which focuses on algorithms that allow programs to be capable of learning and modifying their structure through learning based on a set of inputted data [30].

A machine learning algorithm has three components: Representation, which describes the way knowledge is represented (e.g. Support Vector Machines, Decision Trees, Neural Networks, Model Ensembles, etc.); Evaluation, concerning how programs are evaluated (e.g. accuracy, specificity, sensitivity, squared error, etc.); and Optimization. [31] Machine learning algorithms differ vastly from each other based on the way programs are represented and depending on the method of learning through a particular scope of programs [32].

Being a field between computer science and statistics, and at the core of data science

and AI, machine learning has been a topic of great activity and interest in the field of biomedicine since they provide the possibility of improving the accuracy of disease diagnosis and early detection, whilst improving the process of decision-making with regard to a patient's status [30]. Machine learning provides methods for analyzing various types of datasets and extracting significant relationships and key characteristics in the data by developing models that best describes the given datasets [33].

E. Supervised Learning

Supervised learning, in the context of artificial intelligence (AI) and machine learning, is a type of machine learning in which both the input and desired output data are included in the training dataset [34]. The objectives in supervised learning are to create a model characterizing the class labels in terms of features found in a particular dataset [35]; and to approximate a mapping function such that when new input data is acquired, predicting the corresponding output data for it can be done [36]. Supervised learning may be categorized into regression or classification.

The defining property of supervised learning is the presence of labelled training datasets. Different types of algorithms under supervised learning generate learning models from these training datasets and the models generated are used to classify unlabeled data from new input datasets [37].

F. Classification in Machine Learning

In machine learning, classification is the process of identifying to which class a new input data belongs, based on a model created through learning from a corresponding training dataset of which the class of the included data is known [38]. Classes are sometimes referred to as categories, targets or labels [39].

A classification model aims to draw some generalization from input data values. Machine learning algorithms used in classification either predict categorical classes or classify data based on the training set and the values (class labels) in classifying attributes and uses it in classifying new data. [36] The resulting classifier model from training is used to assign class labels to the testing datasets of which the values of the predictor features are known, while the class label values are unknown [37].

G. MS Data Pre-processing

Mass spectrometry data, in the form of mass spectra, contain huge quantities of m/z and intensity measurements which represent a great number of peaks. In order to efficiently analyze MS data, some data pre-processing steps are required [11][40]. The steps and algorithms used for each are discussed in the subsequent sections.

1. Noise Reduction

Noise Reduction is defined as the removal of mass spectrometry data noise which may be chemical or electronic in origin. This step may be executed through smoothing or baseline reduction.

1..1 Smoothing

Smoothing is a pre-processing step in which data points are averaged with their neighbors in a series of data. The main reason for applying smoothing is to increase signal to noise ratio of a spectrum [11]. One of the most commonly used smoothing techniques is the **Savitzky-Golay Algorithm**. This algorithm involves performing a least squares fit of a small set of consecutive data points to a polynomial and take the calculated central point of the fitted polynomial curve as the new smoothed data point. A set of integers $(A_{-n}, A_{-(n-1)}, \ldots, A_{n-1}, A_n)$ could be derived and used as weighting coefficients to carry out the smoothing operation. The use of these weighting coefficients, known as convo-

lution integers, turns out to be exactly equivalent to fitting the data to a polynomial, as just described and it is computationally more effective and much faster. Therefore, the smoothed data point $(y_k)_s$ by the Savitzky-Golay algorithm is given by the following equation:

$$(y_k)_s = \frac{\sum_{i=-n}^{i=n} A_i(y_{k+i})}{\sum_{i=-n}^{i=n} A_i}$$

1..2 Baseline Reduction

Baseline reduction involves the removal of the baseline slope and offset from a mass spectrum in order to flatten its base profile [11]. Moreover, this pre-processing step is considered to be a linear operation [41].

The baseline reduction step includes the following:

- Set the first and last m/z values of the spectrum as initial reference points
- Subtract the distance of the reference points from the baseline (offset value) from the other intensity values
- Select the troughs of the spectrum as new reference points
- Calculate the average offset value of the troughs and subtract from other intensity values except the initial reference points
- Repeat steps (except first) until the troughs are adjacent to the baseline

2. Data Normalization

Data normalization enables the comparison of different samples since the absolute peak values of different fraction of spectrum could be incomparable. Moreover, it removes sources of systematic variations between spectra due to varying amounts of sample or degradation over time in the sample or even variation in the instrument detector sensitivity [11]. Some techniques for normalization are:

- **Simple Feature Scaling**

  Let x_0 be an initial value, x_f a normalized value, and A_{max} the maximum value of an attribute. Normalization of x_0 to x_f is done using the following equation:

  $$x_f = \frac{x_0}{A_{max}}$$

  Each value is divided by the maximum among all values of a particular feature in a dataset. This makes the values range between zero and one.

- **Min-Max Normalization**

  Min-max normalization involves linear alteration on chosen initial data where the values are normalized within a given range [42]. Let x_0 be an initial value, x_f a normalized value, A_{min} the minimum value of an attribute, and A_{max} the maximum value of an attribute. For mapping a value x_0 of an attribute A from range A_{min}-A_{max} to a new range A_{newmin}-A_{newmax}, the computation for the transformation is given by:

  $$x_f = \left(\frac{x_0 - A_{min}}{A_{max} - A_{min}}\right) * (A_{newmax} - A_{new\_min}) + A_{new\_min}$$

- **Z-Score Normalization**

  Z-score normalization is one of the most commonly used data normalization techniques [43]. Let x_0 be an initial value, x_f a normalized value, μ_A the mean, and σ_A the standard deviation of the values of an attribute, respectively. The value x_f is obtained using the following equation:

  $$x_f = \left(\frac{x_0 - \mu_A}{\sigma_A}\right)$$

3. Data Reduction

This step is crucial for preserving raw data information while performing a dimensional reduction for subsequent processing [11]. The most common technique is **Binning**, which

performs data dimensionality reduction by grouping measured data into "bins" in order to combine a set of continuous values into a single value. The execution of this process involves the following [40]:

- Split mass spectrum into intervals (bins) of m/z values
- Calculate for each interval of m/z values(bin):
  - an aggregate intensity (e.g. the sum of the intensities in the bin)
  - a representative m/z value (e.g. the median of the one with maximum intensity)
- Replace each bin with the calculated representative m/z value

4. Peak Alignment

The peak alignment step is used to find out which peaks among different spectra refer to the same peak, since some mass spectra in a particular MS dataset could correspond to the same chemical compound.

After correcting the retention time shifts in every spectrum, the matching methodology done is as follows [44]:

- Select a reference sample s_R from sample set $S(s_1, s_2, s_3, ..., s_N)$ and denoting the remaining samples as $S' = s_i(i = 1, 2, 3, ..., N - 1)$
- The corresponding peak list of the reference sample is defined as the reference table ($RefTbl$), and the corresponding peak lists of the remaining samples are designated as search tables ($SchTbl$) numbered from $SchTbl_1$ to $SchTbl_{n-1}$. A variable m is initialized $m = 1$.
- Select the m^{th} search table $SchTbl_m$ from the search tables and matching each landmark peak in $SchTbl_m$ to the landmark peaks in $RefTbl$.

H. Support Vector Machines

Support Vector Machines (SVM) is a machine learning algorithm under supervised learning which is applicable for creating models for solving both classification and regression problems [45]. The objective of the SVM algorithm is based on finding the line decision boundary (hyperplane) that provides the largest distance to the closest data points known as support vectors [46]. There are many possible hyperplanes to choose from when separating data points into classes. Consequently, maximizing the margin distance results to a better model since future data values can be classified with higher confidence [47].

Compared to other classification algorithms, Support Vector Machines is less prone to data values which are considered to be outliers since it only regards the values that are nearest to the support vectors or decision boundary [48]. Moreover, SVM is eminently preferred by many due to its capability of producing significant accuracy using less computational power. Even though it can be used for both regression and classification tasks, it is usually used for the latter [47].

Given a labeled training dataset:

$$(x_1, y_1), ..., (x_n, y_n), x_i \in \mathbb{R}^d \text{ and } y_i \in (-1, +1)$$

where x_i is a feature vector representation and y_i is the class label of a training compound i. The hyperplane can be defined as follows:

$$\vec{w}\vec{x}^T + b = 0$$

where w is the weight vector, x is the input feature vector, and b is the bias. The w and b would satisfy the following inequalities for all elements of the training dataset:

$$\vec{w}\vec{x_i}^T + b \geq +1 \text{ if } y_i = 1$$

$$\vec{w}\vec{x_i}^T + b \leq -1 \text{ if } y_i = -1$$

The objective of training an SVM model is to find the $\vec{w}$ and b so that the hyperplane separates the data and maximizes the margin $\frac{1}{||\vec{w}||^2}$.

In the case that the dataset to be used for training is not linearly separable, the "kernel trick" may be used. The "kernel trick" enables the mapping of data to a higher dimensional space where it becomes linearly separable [49].

One the most prominently used kernels is the Gaussian Radial Basis Function (RBF) Kernel. An RBF is a general model used to build a mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$ [50]. When using an RBF, one needs to choose:

- The training function, $\phi(r)$ that will be used. (Some choices: $\phi(r) = e^{\frac{-r^2}{\sigma^2}}$, $\phi(r) = r^3$)
- The model parameters for an RBF
  - The centers c_i, which are also in $\mathbb{R}^n$
  - The number of centers, k (independent of n or m)
  - The weights (or coefficients) $\omega_i, i = 1, 2, ..., k$.

After choosing the centers, the coefficients are obtained using a least squares solution:

- Given p data points in $\mathbb{R}^n$ and k centers in $\mathbb{R}^n$, form the $p \times k$ Euclidean Distance Matrix (EDM) A, such that

$$A_{ij} = dist(x^{(i)}, c^{(j)}) = ||x^{(i)} - c^{(j)}||$$

- Form the transfer matrix Φ: $\Phi = \phi(A)$. If biases are desired, add a column of 1's to the end of Φ, so that it is $p \times (k + 1)$.
- Solve the matrix equation:

$$\Phi \cdot W = Y$$

where W (which has size $(k+1) \times m$) contains the weights of the linear combination, and Y (which has size $p \times m$) contains the desired outputs.

To test the function on new domain points (now with fixed centers and weight matrix W):

- Form the EDM. With $\vec{p}$ new data points, this will be $\vec{p} \times k$.
- Form the transfer matrix Φ, which will be $\vec{p} \times (k+1)$.
- Perform the matrix product ΦW, producing the new output.

I. Evaluation of Classifier

1. K-Fold Cross Validation

In k-fold cross validation, the training set is partitioned into k subsets of equal size. Then, one subset is used as test data using the classification model, after being trained on the remaining $k-1$ subsets. This process is done with k iterations until each subset is used as a testing set [51]. An additional step that could be used for preventing both underfitting or overfitting problems is re-randomization, where there dataset is re-randomized prior to repeating the training and validation processes.

2. Model Evaluation Metrics

2..1 Accuracy

Accuracy is the number of correct predictions outputted divided by the total number of predictions made, converted to percentage [52].

2..2 Recall

Recall measures the ratio of actual positives that are correctly identified as such. It

is also called the true positive rate, the recall, or probability of detection in some fields [53]. As an example, in a medical test used to identify a disease, the recall of the test is the proportion of people who test positive for the disease among those who have it.

2..3 Precision

Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives. False positives are cases the model incorrectly labels as positive that are actually negative [54]. As an example, in a medical test for diagnosing a disease, the precision is the proportion of the true positive over the number of true positives plus the number of false positives.

2..4 F1-Score

F1-Score is the *harmonic mean* between precision and recall ($2 \times \frac{precision \times recall}{precision + recall}$)

IV. Design and Implementation

A. Datasets

The datasets used for creating the SVM Classifier Model consist of mass spectrometry data of samples taken from pancreatic cancer patients and controls who are free from pancreatic cancer. The mass spectrometry data is acquired from the Global Natural Product Social Molecular Networking (GNPS) MS database website and the PRoteomics IDEntifications (PRIDE) Archive, all in *mzml* format. Each particular *mzml* file contains: metadata such as instrumentConfigurationList, aquisitionSettingsList, dataProcessingList, sampleList; and also the spectrum list which contains the spectrum description and the actual spectra data. Validation of file formats and mass spectrometry data pre-processing would be done using pymzML, SciPy, and NumPy.

The dataset to be used for the identification of chemical compounds from mass spectra is acquired from the Chemical Entities of Biological Interest (ChEBI) database. Included are three tables: *compounds*, *names*, and *chemical_data*.

B. System Architecture

The implementation of the system is defined in the following diagrams:

1. Development Flow Diagram

The first step is data collection. The quality and quantity of the data are crucial since it would determine how good the classifier model would be. A higher number of quantity of mass spectrometry data would be beneficial in order to create an accurate model. Moreover, the quality would affect the degree of difficulty the pre-processing would involve.

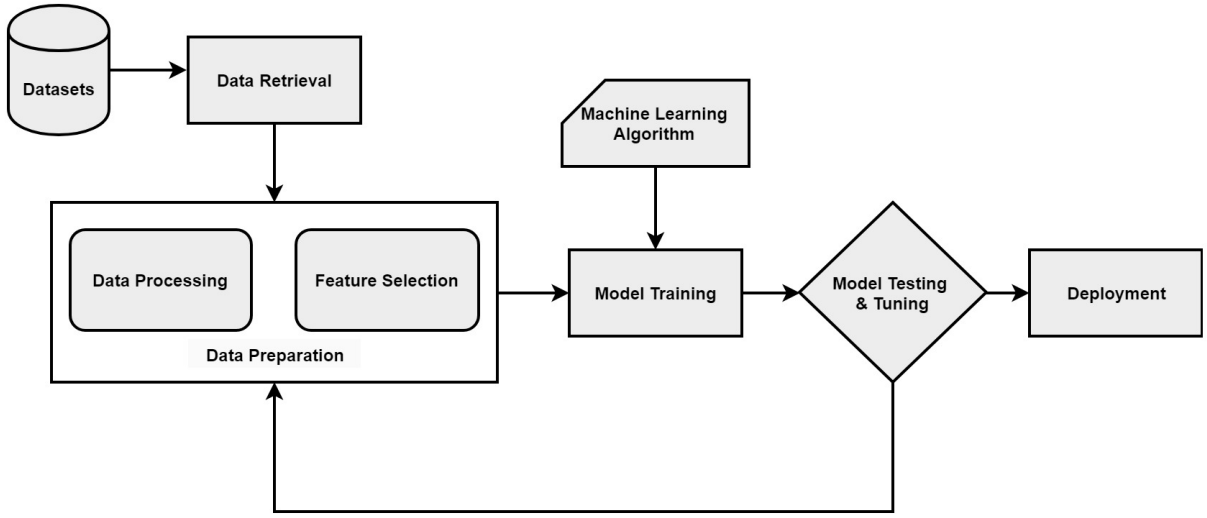


Figure 1: Machine Learning Development Flow Diagram

After gathering enough mass spectrometry data, the preparation (including data processing and feature selection) comes next. The mass spectrometry data pre-processing steps are as described in the preceding chapter. All steps shall be implemented with pymzML, SciPy, and NumPy. With regard to feature selection, since the number of features in a particular mass spectrum is low and all the features of the spectrumList data are considered relevant, all features would be used for the development of the model.

Once the data has been pre-processed, training of model starts. The algorithm to be used is the Support Vector Machines (Radial Basis Function Kernel) algorithm. The original input space would be mapped into a high-dimensional feature space using RBF where it becomes linearly separable. This allows a more efficient classification involving the mass-per-charge, intensity, and retention time values.

Consequently, model testing is done and the percentage values of accuracy, sensitivity and specificity are calculated. Tuning the model parameters is necessary whilst repeating the model training and testing processes iteratively. Subsequently, 10-fold cross validation would be performed.

2. Context Diagram

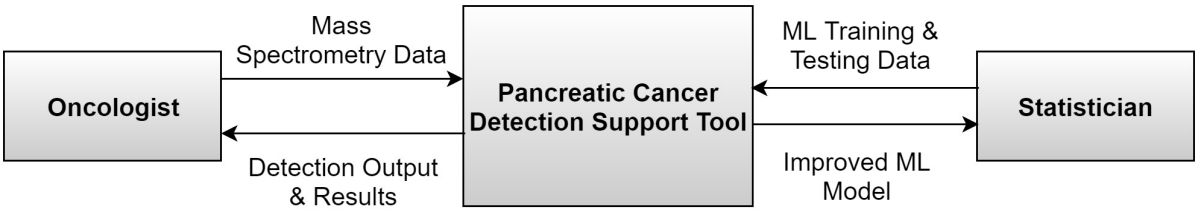


Figure 2: PCDST Context Diagram

The system would be utilized by two types of users: the researchers / oncologists and the statisticians/administrators. The researcher / oncologist would be performing the disease detection by inputting a patient's mass spectrometry data. After data pre-processing, the detection is executed and the results are generated together with the list of most common chemical compounds found in the mass spectrum. On the other hand, the statistician would be in-charge of the machine learning-related work the tool requires. He/she may input additional training datasets in order to improve the current model.

3. Use Case Diagrams

The researcher could perform all the steps in the mass spectrometry data pre-processing sequence. Each step could be selected individually as to provide the user the option whether to perform a particular step or not. The inputted MS data could also be viewed in a visual form (mass spectrum image). Detection of disease and results generation could be done after the preceding steps. Also, a user's manual for the tool would be included.

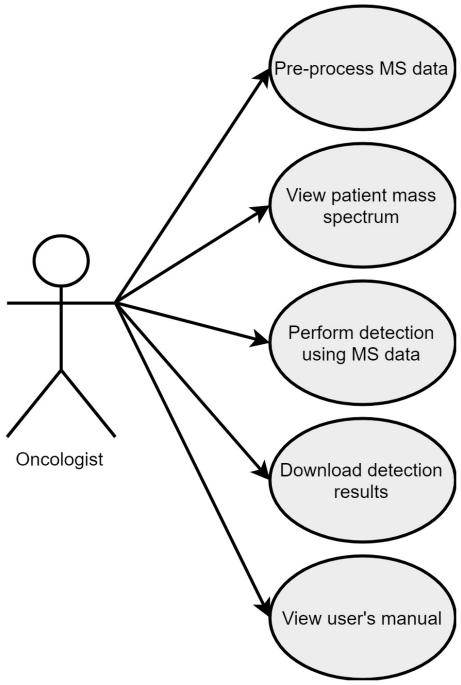


Figure 3: Use Case Diagram of Researcher / Oncologist

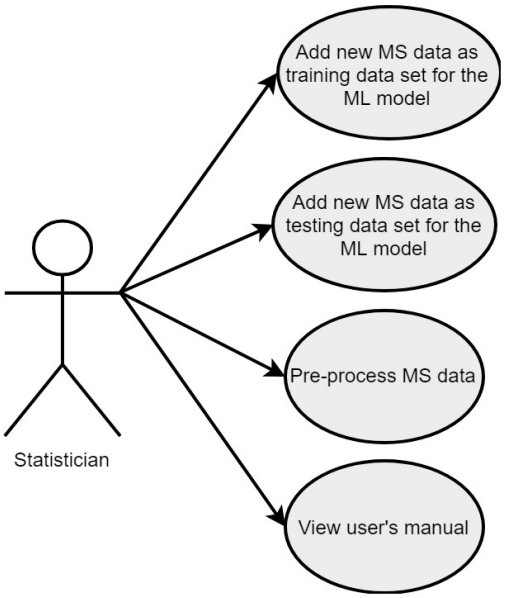


Figure 4: Use Case Diagram of Statistician

The statistician would be able to add both training and testing datasets for the SVM model. Pre-processing of these datasets is recommended. A user's manual for the tool

would be included.

4. Data Flow Diagrams

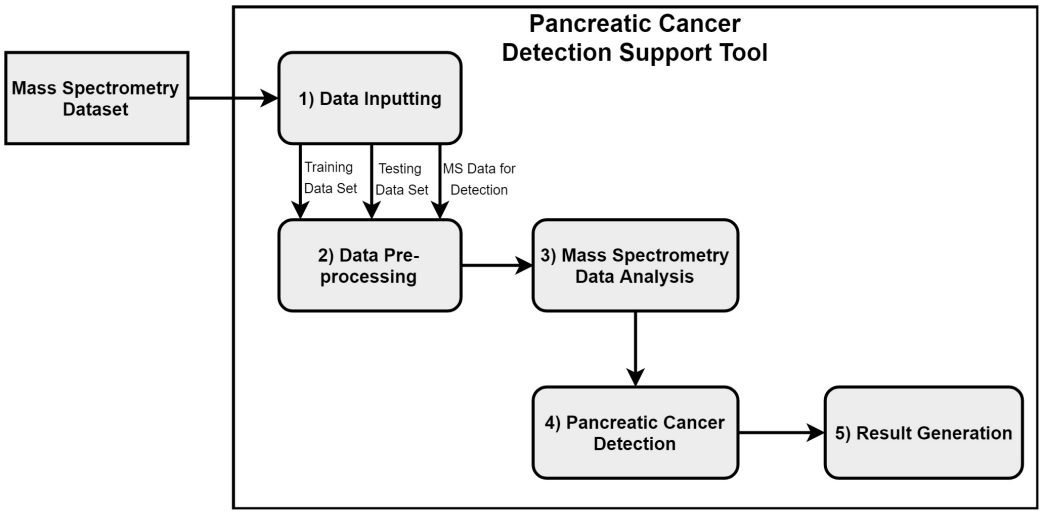


Figure 5: Top-Level Dataflow Diagram of PCDST

For both users, the dataflow starts with the data inputting, where the user uploads the mass spectrum file (*mzml* format required) to the tool. Afterwards, data pre-processing with selected steps is executed. (The succeeding processes would only be performed by the oncologist).

The patient's mass spectrum could be then analyzed in graph form, complemented by the tabulated form of data. Detection is done subsequently, showing the percentage value of likelihood (based on data's distance to decision boundary). Finally, the results are generated and the options for result file download is presented.

The processes in the top-level dataflow diagram is illustrated in Figures 6-8 below:

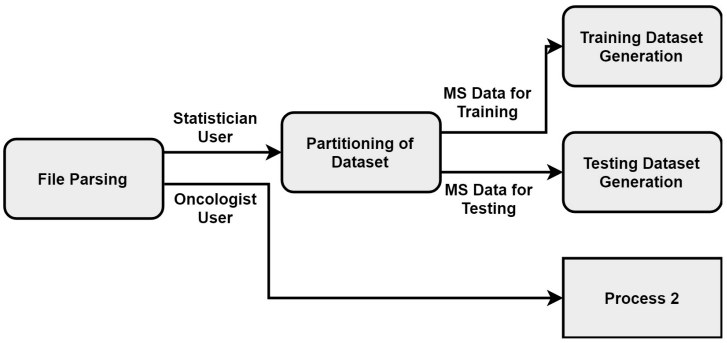


Figure 6: Dataflow Diagram of Process 1

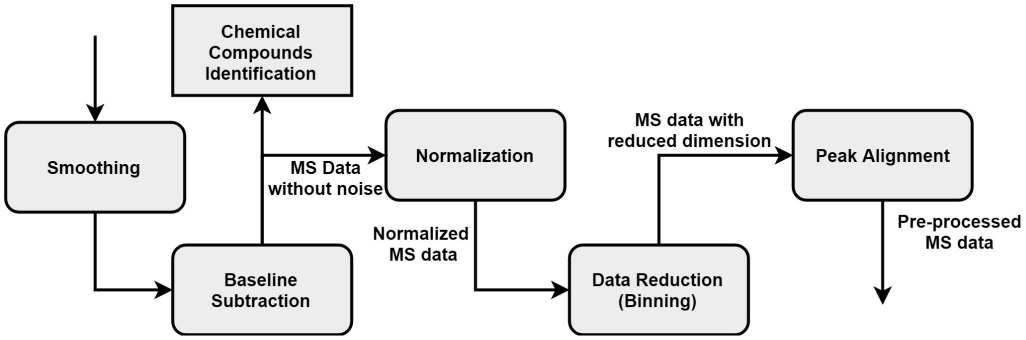


Figure 7: Dataflow Diagram of Process 2

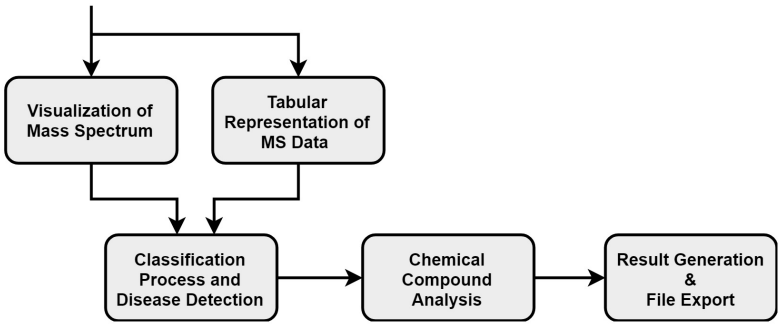


Figure 8: Dataflow Diagram of Processes 3-5

5. Entity Relationship Diagram

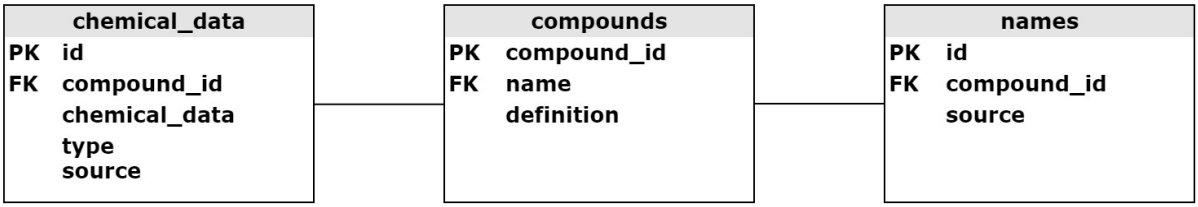


Figure 9: Entity Relationship Diagram of Chemical Compounds Database

The ERD of the chemical compounds database includes three tables: compounds, chemical_data, and names. These tables are utilized for determining the most abundant chemical compounds in a particular patient's mass spectrum, based on the mass-per-charge values with the highest intensities.

C. Technical Architecture

The system would be developed using the following:

- PyCharm IDE 2018.3
- Python 3.7
- NumPy 1.16
- SciPy 1.2
- pymzML 2.0.6
- scikit-learn LibSVM 3.23
- MySQL
- Windows 10 OS

Minimum hardware requirements:

- At least 4 GB free hard-disk space
- At least 8 GB RAM
- Minimum processor speed of 2.6 GHz

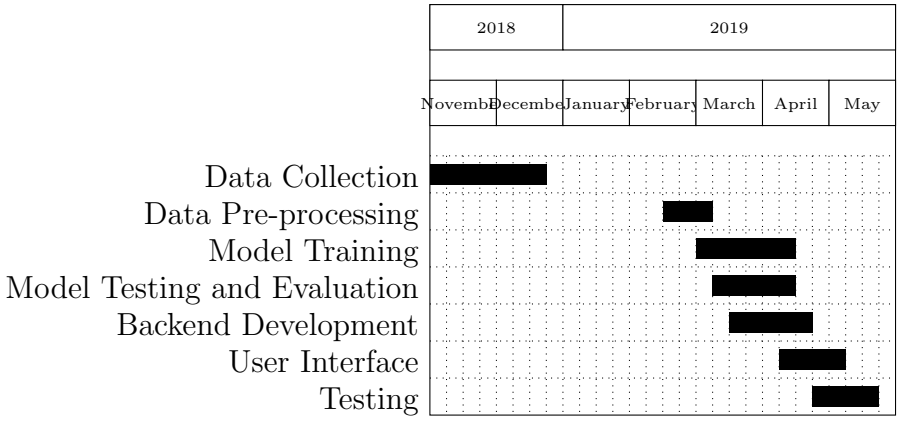
V. Expected Output

This study is expected to have the following output:

Pancreatic Cancer Diagnosis Tool

- Main Window
  - Oncologist Panel
    - Control Panel for Data Input and Pre-processing
    - Mass Spectrometry Data Analysis Tab
    - Analysis Results Window
    - Results Export Tool
  - Statistician Panel
    - Control Panel for Data Input and Pre-processing
    - Model Training Panel
    - Model Evaluation Results Tab
  - User's Manual

VI. Schedule of Activities



VII. Bibliography

[1] W. C. R. Fund, "Pancreatic cancer statistics," 2018. [Online]. Available: https://www.wcrf.org/dietandcancer/cancer-trends/pancreatic-cancer-statistics?fbclid=IwAR3R4-G3KzUh6sXmDCbWXIchsGYVQvXGdpqUgjxGZ90driz4p8NlmteTIXA

[2] M. Ilic and I. Ilic, "Epidemiology of pancreatic cancer," 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/ONLINEs/PMC5124974/

[3] A. C. Society, "Pancreatic cancer risk factors," 2016. [Online]. Available: https://www.cancer.org/cancer/pancreatic-cancer/causes-risks-prevention/risk-factors.html

[4] L. Martin, "An overview of pancreatic cancer," 2017. [Online]. Available: https://www.webmd.com/cancer/pancreatic-cancer/digestive-diseases-pancreatic-cancer

[5] A. Ashcroft, "An introduction to mass spectrometry," n.d. [Online]. Available: http://www.astbury.leeds.ac.uk/facil/MStut/mstutorial.htm

[6] O. Mamer, "Medical applications of mass spectrometry," 2017. [Online]. Available: https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/mass-spectrometry

[7] U. Pancreatic Cancer Organization, "How is pancreatic cancer diagnosed?" 2018. [Online]. Available: https://www.pancreaticcancer.org.uk/diagnosis

[8] A. C. Society, "Can pancreatic cancer be found early?" 2016. [Online]. Available: https://www.cancer.org/cancer/pancreatic-cancer/detection-diagnosis-staging/detection.html

[9] A. Swan, A. Mobasheri, D. Allaway, S. Liddell, and J. Bacardit, "Application of machine learning to proteomics data: Classification and biomarker identification in postgenomics biology," *OMICS: A Journal of Integrative Biology*, p. 595–610, 2013, doi: 10.1089/omi.2013.0017.

[10] J. Rajapakse, K. Duana, and W. K. Yeo, "Proteomic cancer classification with mass spectrometry data," *American Journal of PharmacoGenomics*, p. 281–292, 2005, doi: 10.2165/00129785-200505050-00001.

[11] K. Coombes *et al.*, "Understanding the characteristics of mass spectrometry data through the use of simulation," 2007. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/ONLINEs/PMC2657656/

[12] C. Smith, E. Want, G. O'Maille, R. Abagyan, and G. Siuzdak, "Xcms: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification," p. 779–787, 2006, doi: 10.1021/ac051437y.

[13] Datta, "Feature selection and machine learning with mass spectrometry data. methods in molecular biology," p. 237–262, 2013, doi: 10.1007/978-1-62703-392-3_10.

[14] S. Ahmed, M. Zhang, and L. Peng, "Feature selection and classification of high dimensional mass spectrometry data: A genetic programming approach," p. 43–55, 2013, doi: 10.1007/978-3-642-37189-9_5.

[15] K.Kim, S.Ahn, J.Lim, B.C.Yoo, J.H.Hwang, and W.Jang, "Detection of pancreatic cancer biomarkers using mass spectrometry," p. 43–55, 2013, doi: 10.1007/978-3-642-37189-9_5.

[16] T. Nguyen, S. Nahavandi, D. Creighton, and A. Khosravi, "Mass spectrometry cancer data classification using wavelets and genetic algorithm," p. 3879–3886, 2015, doi: 10.1016/j.febslet.2015.11.019.

[17] G. Ge and G. Wong, "Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles," p. 275, 2008, doi: 10.1186/1471-2105-9-275.

[18] X. Zhang, X. Lu, Q. Shi, X. Xu, H. Leung, L. Harris, and W. Wong, "Recursive svm feature selection and sample classification for mass-spectrometry and microarray data," p. 3879–3886, 2006, doi: 10.1186/1471-2105-7-197.

[19] P. Pham, L. Yu, M. Nguyen, and N. Nguyen, "Fast cancer classification based on mass spectrometry analysis in robust stationary wavelet domain," p. 189–199, 2011, doi: 10.1007/978-94-007-2598-0_21.

[20] A. McNaught, A. Wilkinson, and A. Jenkins, *Compendium of Chemical Terminology*. IUPAC, 1997, online corrected version: (2006–) "mass spectrum".

[21] ——, *Compendium of Chemical Terminology*. IUPAC, 1997, online corrected version: (2006–) "Ionization".

[22] K. Strimbu and J. Tavel, "What are biomarkers?" 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/ONLINEs/PMC3078627/

[23] Mayoclinic, "Pancreatic cancer," 2018. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/pancreatic-cancer/symptoms-causes/syc-20355421

[24] A. S. of Clinical Oncology (ASCO), "Pancreatic cancer: Symptoms and signs," 2019. [Online]. Available: https://www.cancer.net/cancer-types/pancreatic-cancer/symptoms-and-signs

[25] P. C. A. Network, "What are pancreatic cancer biomarkers?" 2018. [Online]. Available: https://www.pancan.org/news/what-are-pancreatic-cancer-biomarkers/

[26] Mayoclinic, "Pancreatic cancer," 2018. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/pancreatic-cancer/diagnosis-treatment/drc-20355427

[27] WebMD, "Pancreatic cancer treatments by stage," 2018. [Online]. Available: https://www.pancan.org/news/what-are-pancreatic-cancer-biomarkers/

[28] P. Urban, "Quantitative mass spectrometry: an overview," 2016, doi: 10.1098/rsta.2015.0382.

[29] C. Wickremasinghe, "How mass spectrometry has changed the cancer diagnostics game," 2018. [Online]. Available: https://www.technologynetworks.com/diagnostics/ONLINEs/how-mass-spectrometry-has-changed-the-cancer-diagnostics-game-300408

[30] P. Sajda, "Machine learning for detection and diagnosis of disease," *Annual Review of Biomedical Engineering*, 2006, doi: 10.1146/annurev.bioeng.8.061505.095802.

[31] J. Brownlee, "Basic concepts in machine learning," 2015. [Online]. Available: https://machinelearningmastery.com/basic-concepts-in-machine-learning/

[32] M. Jordan and T. Mitchell, "Machine learning: Trends, perspectives, and prospects," p. 255–260, 2015, doi: 10.1126/science.aaa8415.

[33] N. Maity and S. Das, "Machine learning for improved diagnosis and prognosis in healthcare," 2017, doi: 10.1109/aero.2017.7943950.

[34] M. Haughn, "Supervised learning," 2016. [Online]. Available: https://searchenterpriseai.techtarget.com/definition/supervised-learning

[35] J. Akinsola, "Supervised machine learning algorithms: Classification and comparison," 2017, doi: 10.14445/22312803/IJCTT-V48P126.

[36] S. Shukla, "Regression and classification - supervised machine learning," 2017. [Online]. Available: https://www.geeksforgeeks.org/regression-classification-supervised-machine-learning/

[37] P. Cunningham, M. Cord, and S. Delany, "Supervised learning. cognitive technologies," p. 21–49, n.d., doi: 10.1007/978-3-540-75171-7_2.

[38] V. Trinadh, "Machine learning - overview of classification problems," 2017. [Online]. Available: https://www.linkedin.com/pulse/machine-learning-overview-classification-problems-trinadh-venna

[39] S. Asiri, "Machine learning classifiers," 2018. [Online]. Available: https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623

[40] M. Cannataro *et al.*, "Preprocessing, management, and analysis of mass spectrometry proteomics data," 2007. [Online]. Available: https://pdfs.semanticscholar.org/3f87/fef3fc2e447d458bb229f0ce8fcdd3a458f2.pdf

[41] S.J.Luck, "An introduction to the event-related potential technique, second edition," 2014. [Online]. Available: https://erpinfo.org/order-of-steps

[42] C.Saranya and G.Manikandan, "A study on normalization techniques for privacy preserving data mining," 2013. [Online]. Available: https://pdfs.semanticscholar.org/35a8/7b51f7441a87adee91e12eb4d22cd2565556.pdf

[43] A.B.Khalifa, S.Gazzah, and N.E.B.Amara, "Adaptive score normalization: A novel approach for multimodal biometric systems," 2013. [Online]. Available: https://waset.org/publications/17136/adaptive-score-normalization-a-novel-approach-for-multimodal-biometric-systems

[44] B.Wang, A.Fang, J.Heim, B.Bogdanov, S.Pugh, M.Libardoni, and X.Zhang, "Disco: Distance and spectrum correlation optimization alignment for two dimensional gas chromatography time-of-flight mass spectrometry-based metabolomics," 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2891529/

[45] A. Shmilovici, "Support vector machines," 2010, doi: 10.1007/978-0-387-09823-4_12.

[46] OpenCV, "Introduction to support vector machines," 2018. [Online]. Available: https://docs.opencv.org/2.4/doc/tutorials/ml/introduction\textunderscoreto\textunderscoresvm/introduction\textunderscoreto\textunderscoresvm.html

[47] R. Gandhi, "Support vector machine—introduction to machine learning algorithms," 2018. [Online]. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[48] M. Swamynathan, "Step 3 – fundamentals of machine learning. mastering machine learning with python in six steps," p. 117–208, 2017, doi: 10.1007/978-1-4842-2866-1_3.

[49] H.Kandan, "Understanding the kernel trick," 2017. [Online]. Available: https://towardsdatascience.com/understanding-the-kernel-trick-e0bc6112ef78

[50] "Rbf summary," n.d. [Online]. Available: http://people.whitman.edu/~hundledr/courses/M350S04/rbfsummary.pdf

[51] J. Brownlee, "A gentle introduction to k-fold cross-validation," 2018. [Online]. Available: https://machinelearningmastery.com/k-fold-cross-validation/

[52] ——, "Classification accuracy is not enough: More performance measures you can use," 2014. [Online]. Available: https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/

[53] D. Altman and J. Bland, "Diagnostic tests. 1: Sensitivity and specificity," 1994, doi: 10.1136/bmj.308.6943.1552.

[54] W. Koehrsen, "Beyond accuracy: Precision and recall," 2018. [Online]. Available: https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c