SciMedCentral®

JSM Computer Science & Engineering

Research Article

K-NN Classification of Mass Spectra Data for Diagnosing Alzheimer's Disease

Destiny EO Anyaiwe[1], **George D. Wilson**[2], **Timothy J. Geddes**[2], **and Gautam B. Singh**[1]*

[1]*Department of Computer Science and Engineering, Oakland University, USA*
[2]*William Beaumont Hospital, Royal Oak, USA*

***Corresponding author**
Gautam Singh, Department of Computer Science and Engineering, Oakland University, Rochester, MI, 48309, USA; Tel: 1248-370-2129; Email: singh@oakland.edu

Submitted: 01 October 2016

Accepted: 15 November 2016

Published: 17 November 2016

Index Terms

- Feature matrix
- Point-to-point discrimination
- Protein biomarkers
- Jackknife

Abstract

Diverse algorithms and methods are needed to answer the ever increasing need of adequately harnessing Mass Spectrometer generated data. The unique nature and structure of this data, requires a high level of expertise and rigorous algorithms to harness its full benefits. The methodology of this study discusses feature selection based on direct observations of variables and their inter-relationships, Jackknife technique for data sampling, matrix to vector decomposition and successfully classifies Alzheimer's disease patients into three disease stages; age-matched controls without any evidence of dementia, patients with mild cognitive impairment and patients with clinical symptoms of Alzheimer's disease (AD). Our model extends the use and principle of K-nearest neighbor (KNN) algorithm and also presents a modification of Euclidean distance formula. Hitherto, there exists no clinical diagnostic tool for AD, in lieu of this, patient cognitive abilities are clinically followed-up over a period of time (may be months) to make a diagnosis. This practice usually leads to inconclusive diagnosis and results obtained from it are not generalizable. This study, provides a platform for immediate classification and correctly indicates test data sets predisposed to AD with 75% accuracy (giving a probability of 0.13 for committing type II error) without collaborating clinical records.

ABBREVIATIONS

Con: Without Evidence of Ad; Mci: Mild cognitive impairment; Tad: Diagnosed of Alzheimer's disease; Sla: Supervised learning algorithms

INTRODUCTION

Proteomics task of discovering and identifying the set of proteins expressed by individual cells with regards to time and other biochemical conditions has witnessed tremendous achievements in recent times due to the invention and introduction of high throughput assay processes like Protein Chip Mass pectrometer–Surface Enhanced Laser Desorption/Ionization (SELDI) time-off light laboratory technique, in that, protein analysis are more readily accessible and available in real time, but in practice, harnessing the output of SELDI experiment involves onerous tasks and demands investigators with high level of expertise.

SELDI provides detailed analysis of the analytes with accurate results; it entails the ionization of analyte (protein) samples by subjecting an analyte to laser energy bombardment. Upon this, elucidated ions are separated based on their mass-to-charge ratio. The feature of these separated ions are recorded and presented as mass spectra showing the relative abundance of 'bio-chemical compounds' contained in the analyzed sample. Each ion in the abundance spectra (assay results) is typically categorized by the following properties; the mass-to-charge ratio (m/z), time-of-flight (TOF), intensity (TOF Intensity), Substance mass, ion charge, ionmass, signal-to-noise ratio and peak type.

For a protein source (e.g. serum, urine, saliva) analytes, SELDI generates hundreds of peptide peaks which are further investigated with respect to the investigator's objectives. The investigation of peptide peaks usually begins with detecting the set of peaks that are 'differentially expressed' in the mass spectra after baseline subtraction has been done using statistical methods or thresholding. Usually, for each protein source analyzed ,a set of differentially expressed peaks (tens to hundreds, depending on the laser energy bombardment level and the type of Protein Chip used) are chosen; these represents the result of the assay process, a collection of which is our raw data.

To identify the protein or peptide of interest, the molecular weights and chemical properties of ions contained in the SELDI raw data, i.e. which chemical surface it binds to preferentially on the Protein Chip, is matched with public databases. Definitive identification of the peak is then carried out using other mass spectrometry methods.

Questions about detecting the bio-chemical changes in cells or tissues that are capable of causing post-translational modifications of proteins or change in protein's structural information, etc are answered using identified peaks .Other uses of SELDI data is in the area of determining molecular formulas, protein curating and identification, and protein bio-marker discovery [1], personalized medicine, drug design and drug production [2,3].

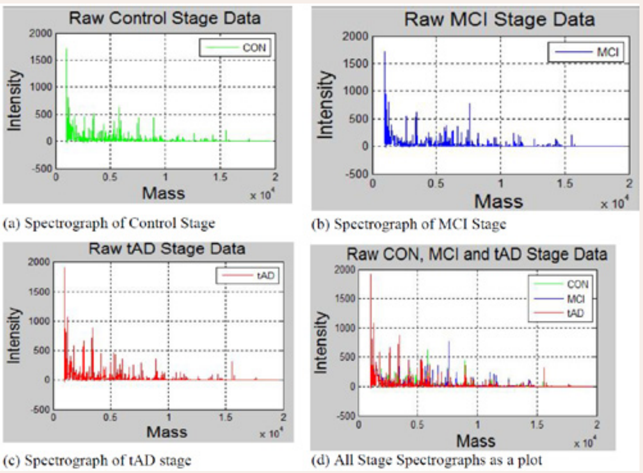
In the US, from 2000 to 2013 while deaths from other diseases declined significantly, that of Alzheimer's disease (AD) increased by 71%. AD is one of the most expensive health conditions to treat in the world today. The estimated cost of care for AD in the US exceeded $214 billion in 2014, with nearly one in every five dollars spent by Medicare on dementia. Future cost estimate from the United States Alzheimer's Association [4], predicts that by 2050 the disease will cost $1.2 trillion annually. The disease currently affects five million people in the US, and expected to grow to 16 million by 2050; afflicting one in nine people over the age of 65, and one in three people over the age of 85.

The clinical practice of diagnosing AD today consists of patients follow ups; patient cognitive abilities (like memory) are tested over a period of time. The practice is time consuming, the follow up can be for many months and may not be conclusive, mild cognitive impairment (MCI) cases may degenerate to full blown dementia (tAD)during this period thereby causing severe and irreversible brain damage to the dementia patient. Additionally, the results or clinical notes achieved by patients follow ups are not generalizable.

Despite the progresses in identifying and discovering several protein bio-markers for Alzheimer's disease, the story is yet to be palatable for its patients and care givers due to lack of clinical diagnostic tools. Consequently, the need for studies such as this.

Harnessing SELDI data involves the application of diverse rigorous statistical or machine learning algorithms towards the investigator's goals. This study, goes beyond protein curating and protein bio-marker identification (which in most cases, are the clinical/laboratory objectives of detail study of SELDI data), to the building of classification model using Mass Spectrometer-SELDI *saliva* data. The end goal is to close the gap between identified bio-marker and the diagnoses of Alzheimer's disease using an extended principle of K-Nearest Neighbor (KNN) Algorithm.

In the course of this study, we considered each output of MS analysis (which are basically, collections of ions and their features (peaks) that were differentially expressed) as matrices, see equation (1). Generally, pictorial view of a data-set may help identify unique patterns, consider (Figure 1), which is a display four different plots; sub-figures (1a), (1b) and (1c) respectively represents plots of the data that represents the three stages of AD; CON, MCI and tAD, and sub-figure (1d) is a display of sub-figures (1a), (1b) and (1c) on same plot. Sub-figure (1d) puts the subject of this study into a clear perspective; the task to achieving or inducing a separation line on elements of the data-sets. From sub-figure (1d), it is easy to see that the location of peaks for all three stages overlaps (i.e, given a mass value there exists an intensity value for all three stages) with no *cluster*, no *regression* or a *discriminative pattern*.



(a) Spectrograph of Control Stage
(b) Spectrograph of MCI Stage
(c) Spectrograph of tAD stage
(d) All Stage Spectrographs as a plot

Figure 1 Spectrograph of Stage Data-sets.

The scenario that every data point of our data-set is a matrix further complicates the aim of this study in that, application of traditional supervised learning algorithms (SLA) fails. Traditional SLA does point-to-point discrimination of points/feature vectors in lieu of matrix-to-matrix discrimination. It is therefore, imperative to first overcome these challenges if positive advancement is to be made in putting SELDI data into additional uses.

KNN is a non-parametric algorithm used for supervised learning, its discrimination of points of a data-setentails casting a net around points of the test data-set. The size of the net depends on the number (k) of points of the train data-set the investigator allows to be the neighbors of a test data point; the discrimination process at the end is by vote and favors the label that is mostly represented in the net. Different distance metrics can be adopted by KNN depending on how the metric function chosen fits the data-set structure. In this study, we extend the principle of KNN and modify the Euclidean distance function in other to apply them to our data-set.

The next section gives a literature review of works done using KNN principle and the description of our methodology. Discussion of our results and observations is given in Section 3 and possible areas of future works and conclusion is highlighted in Section 4.

MATERIALS AND METHODS

Literature

KNN has gained its place in diverse areas of studies; sciences, business, medicine, and in solving on-line and social networks, speech, text and image recognition problems. Its generalization principle uses an appropriate distance function to induce measures on the locations of instances of the train data-set from a test data-point; study results have shown it is most adept for data-sets with 3-to-4 classes [5].

The efficacy of machine learning algorithms in solving any supervised or unsupervised learning problem greatly depends on the algorithm's approach and the data structure under study [5,6], KNN is a non-parametric algorithm, easy to implement, modify and extend.

In general, it is advantageous to conceptualize individual objects (e.g. genotypes) as elements existing in a multidimensional space, this way, geometric classification techniques can be applied to create homogeneous groups by building data from the structure of correlated groups in the multidimensional space [7]. Data structure plays vital role in solving classification problems, sometimes, it renders the data insensitive for analysis by either hiding or camouflaging important details in the data-set. Different approaches exists that can be applied to select objects (e.g. gene) from a genomic data-sets, Leping et al., in [8] explored KNN with Genetic Algorithms as an approach for the generation of predictive gene subsets.

Application of dimensionality reduction or feature extractions to a data-set reduces the number of features in the data-set which in turn enhances the usage of the resulting data-set while eliminating the possibilities of over fitting problems. In [9], multi-labeling based on identifying the KNNs of training set in instances of test set was presented, it further showed how such exercise can be used to predict yeast gene functionality, assign labels to unseen images in natural scene classification problems and solve web page automated categorization problems, similar idea was presented in [10] for image recognition.

Similar to DNA sequence alignment, structural proteomics was studied in [11]. The study achieved grouping and predicting of new proteins based on structure alignments of the distance matrices obtained by 2D representation of protein's tertiary structures.

Sundry studies about phylogenetic tree constructions, node connections in social and biological network systems utilize different forms of distance functions [12,13]. The results of such studies can be extended for classification or predicting purposes if supplemented with the generalization principles of KNN.

This study utilized, Jackknife sampling procedure to constitute elements of the training and corresponding test data-sets. The importance and reasons as to why and when Jackknife technique can be used were presented by [14]. The method was applied for feature selection and classification in [15].

Methodology

The data-set used for this study was obtained from the Bio Bank of Beaumont Reference Laboratory and was the output of a Surface Enhanced Laser Desorption/Ionization time-of-flight (SELDI-TOF) discovery proteomics laboratory experiment carried out on saliva. The experiment was designed to assess differential protein expression sin saliva donor samples for the purpose of identifying protein biomarkers for Alzheimer's disease (AD). Three populations of patients were studied consisting of age-matched controls without any evidence of dementia (CON), patients with mild cognitive impairment (MCI) and patients with clinical symptoms of Alzheimer's disease (tAD).

Also of note is that, having so many (tens, sometimes hundreds of) observations in an experimental result as inherent in high-throughput assay procedures like SELDI-TOF and MALDI-TOF analysis, throws-in another form of problem to feature selection, pattern recognition and building of classification models and tools. This is because, traditional Supervised/Unsupervised machine learning algorithms accepts feature vectors as inputs and discriminate data points on a point-to-point basis, i.e. given a data set and based on the parameters of a chosen algorithm, every instance of the test data-set is examined and subsequently labeled or added to a cluster group depending on the type of problem being solved. Whereas, in particular, SELDI output data is made up of matrix data points.

We present a basic systematic approach for feature selections and transformed matrices contained in the data-set to collections of feature vectors. A distance metric called exponential Euclidean distance function was also introduced. The classification model described in this study; classifies and predict test samples into the 3stages of Alzheimer's disease using K-nearest neighbor (KNN) classifier. This was achieved by assigning a test data to the stage with the highest number of k-minimum distance hits in each iteration for k = 1 and k = 5.

Data organization: The 'uniqueness' of the raw data-set is as a result of the structure of each data-point (SELDI analysis result) it contains. Matrix (R) represents the 179 differentially expressed peaks selected as the result for each saliva sample analyzed. Every matrix R has two types of attributes; Numerical attributes (M/Z, ToF, ToF Intensity, Substance Mass, Charge, ion Mass and Signal to Noise) and a categorical attribute; Peak Type with the values, first pass, second pass and estimated peak types.

$$R_k^S = \begin{bmatrix} m_1 & T_1 & I_1 & S_1 & C_1 & M_1 & N_1 & P \\ m_2 & T_2 & I_2 & S_2 & C_2 & M_2 & N_2 & P \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \; \vdots \\ m_n & T_n & I_n & S_n & C_n & M_n & N_n & P \\ & & & & & & & \end{bmatrix} \quad (1)$$

Where[1] k and S are additional parameters we took the benefit to introduce $k = \{1, 2, \ldots, 20\}$, indexes the total number of results in each disease stage *(S)*.

In each R, there are $n = 1, \ldots, {}^{R}79$ number of rows of observed ions and elements of are arranged in an ascending order that relates to the size of m/z values; i.e. $m_1 < m_2 < \ldots < m_{179}$.

Mass spectrometer calibrates the features values with different scales (as $V1$ indicates), applying principle component analysis (PCA) led to loss of sensitivity of the data due to the features selected. The following observations $[V1 - V5]$ was used to identify the interrelationships that exists between columns of R and was also used to achieve pre-processing of R the data as well as feature selection.

<?> $^{1}m_{n}$ is m/z (or molecular mass), T_n stands for time-of-flight (TOF), I_n denotes TOF Intensity, S_n is Substance mass, C_n for an ion charge, M_n n mass, *N* for signal-to-noise and *I* implies peak type

$V1:$ TOF values are small $(T_i \times 10^{-5})$ and approximates to 0.0000 (at 4 decimal places). Also, for all ion in *R*. The entry values for ion charge and ion Mass is 1 $(C_i = M_i = 1)$. Leaving out these features will only cause a uniform perturbation (if any) to the data-set.

$\Rightarrow \forall i \exists \alpha s.t. \alpha_i = m_i \times T_i$ As the mass (m/z) values increases in size, TOF Intensity (I_i) values are relatively decreasing, i.e.

$$mass = \frac{1}{TOFIntensity}$$

$$\Rightarrow \forall i \exists \alpha s.t. \alpha_i = m_i \times T_i$$

Similar relation also exists between substance Mass and TOF Intensity but the value of α is not constant across rows of *R*, moreover, substance Mass and m/z are related as expressed by (V_4), thus, both cannot be used in a model.

V3: The parameters of Peak Type are; First Pass, Second Pass and estimated, these parameters are used to reference when, during the analysis process the Mass Spectrometer machine recorded such peaks. Some ions are more stable and travels through the Mass Spectrometer machine without further fragmentation, the peak of such ions are registered as first pass while the peak of ions that are results of further fragmentation are recorded as second pass peaks. Estimated peaks are average peaks assigned by the researcher. If anion's peak is not registered in an analysis result but such ion has a peak in the pool of results, the average of available peaks is evaluated, assigned and remarked as estimated for the missing peak. The implication of this is discussed in future works.

V4: V4: $m_i = S_i + C_i$; for any *ion$_i$*, the sum of its substance Mass and Charge equals its molecular mass value.

V5: Signal-to-Noise values are higher for First Pass peaks and relatively equal and smaller for Estimated and Second Pass peaks (same thought as in V3).

Sequel to these observations, the matrix (R_k^S) was reduced to a *2-by-179* matrix (p_k^S) shown below, having only the m/z (m_n^k) and TOF Intensity (I_n^k) features.

$$p_1^C = \begin{bmatrix} m_1^1 & I_1^1 \\ m_2^1 & I_2^1 \\ \vdots & \vdots \\ m_n^1 & I_n^1 \\ & \end{bmatrix} p_5^M = \begin{bmatrix} m_1^5 & I_1^5 \\ m_2^5 & I_2^5 \\ \vdots & \vdots \\ m_n^5 & I_n^5 \\ & \end{bmatrix} \qquad (3)$$

Above is the snapshots of two (p_k^S) matrices, the matrix on the left is the first as contained in the CON dataset while the matrix on the right is the fifth in the MCI stage dataset, *k=1,...,20,* is the numbering for the 20 data points in each stage (*S*) and *n* in (m_n^k or I_n^k) is row-wise numbering of the ions in each matrix (*p*). Going forward, we shallsimply refer to m/z as mass, TOFIntensity as intensity, and an ion as a peak defined by the pair (*mass; intensity*).

The Data-Sets: The population used for the SELDI discovery proteomics was sub-typed into CON, MCI, and tAD stages based on disease severity and each stage has 20 Spectra results, with each data point (p) having 179 rows (or peaks).

To proceed, we recall some basic notes about matrices and vectors;

1. A row matrix is a matrix that has only one row.
2. A column matrix is a matrix that has only one column.
3. A matrix with only one row or one column is called a vector.

More elaborate definitions and proofs were given by Wangmeng et al., in [10]. Based on the above notes, the feature matrices (*p*) were transformed to vectors by simply dropping the notion of matrix and treating each row in (p) as individual row vectors, as shown by (Figure 2). Thus, each vector holds unique information about a unique peak including a label to denote the stage the peak belongs to. The principle of Jackknifing was then used to generate the train and corresponding test data-sets.

Definition 2.1: Jackknife Procedure: This procedure is a re-sampling without replacement technique used to correct bias or create confidence limits for estimators and advisable in scenarios were there exist no statistical or biological models to test new research results with. Given a sample (*X*) of size *N* a *delete-d* Jackknife samples is obtained by selecting and deleting *'d'*-number of observations from the sample. For each Jackknife sample, parameters are estimated and tested on the deleted sample, then the final Jackknife estimate is achieved by taking the aggregate of the *'d'* estimates thus generated. For instance, a *delete-1* Jackknife sample will look like;

$$\mathrm{X_a} = \mathrm{X_b}, \mathrm{X_c}, \ldots, \mathrm{X_n} \qquad (4)$$

Xa

is used as the test data while terms on the right hand side of Eq.4 constituted the elements of the train data-set. Our raw data-set has 20 data-points in each stage, thus, 20 Jackknife training data-sets (adopting the delete-1 Jackknife procedure) for each stage (60 in all), was generated. For iteration, a training data-set is learned and consequently used to test the corresponding test data-set.

Definition 2.2: Exponential Euclidean Distance: In general,

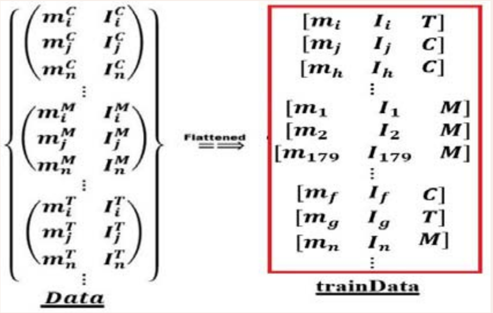


Figure 2 Matrix Data to Row Vectors (Left) is the preprocessed data set made up of matrices. (Right) is the flattened data set made up of vectors; the alphabets C, T, &M stands for control, mci and tAD respectively.

there exists three cases that may exist between any two peaks, these are; 1) their molecular mass values are (approximately) equal but their intensity values differs, 2) they both have unequal molecular mass and intensity values and 3) they have unequal mass values but equal intensity values. case 1 is most profitable, it indicates measuring and comparing the abundance level of peaks provided they have equal mass values (i.e, both peaks must exists in the same horizontal location), case 3 measures equal peak intensity's irrespective of their mass values or horizontal locations (this case is not informative, it is just comparing the obvious; molecular mass values) while case 2 exists as an alternative for the model to rely on if case 1 fails.

The concept of KNN is based on minimum values so concentrating on case 1, the popular Euclidean distance function Eq.5;

$$d(a,b)=\sqrt{(m_a-m_b)^2+(I_a-I_b)^2} \qquad (5)$$

defined between two vectors *a* and *b* is not directly applicable here, since we need a formula that is biased towards intensity values. In particular, the terms (m_a-m_b) and (I_a-I_b) of Eq.5 are evaluated on the same scale. We then introduced and established Eq.6, called the *exponential* Euclidean distance function and defined the distance between two vectors *a* and *b* by

$$dist_{(a,b)}=\sqrt{\left(e^{(m_a-m_b)^2}-1\right)^2+(I_a-I_b)^2} \qquad (6)$$

Where *m* and *I* represents the mass and intensity of row vectors *a* and *b* respectively. By Eq.6, the term $\left(e^{(m_a-m_b)^2}-1\right)$ evaluates to zero if $m_a=m_b$, they by, laying emphases on the other hand the value is (I_a-I_b), exponentially magnified even when the difference between m_a and m_b is very small, in lieu of adding some *'small'* value that might have resulted from (m_a-m_b) if Eq.5 was used. These cases is further explained with Figure (3).

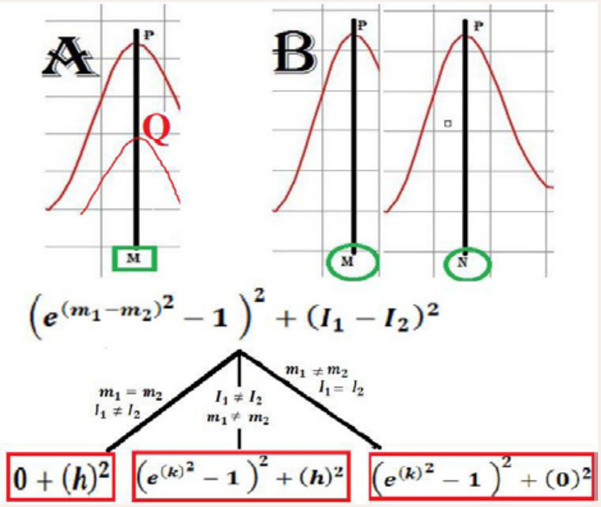

A
P
Q
M
B
P
M
P
N
$\left(e^{(m_1-m_2)^2}-1\right)^2+(I_1-I_2)^2$
$m_1=m_2$
$I_1\neq I_2$
$I_1\neq I_2$
$m_1\neq m_2$
$m_1\neq m_2$
$I_1=I_2$
$0+(h)^2$
$\left(e^{(k)^2}-1\right)^2+(h)^2$
$\left(e^{(k)^2}-1\right)^2+(0)^2$

Figure 2 Matrix Data to Row Vectors (Left) is the preprocessed data set made up of matrices. (Right) is the flattened data set made up of vectors; the alphabets C, T, &M stands for control, mci and tAD respectively.

Definition 2.3: Metric Function: A metric (*d*) in $R^{pcol\times prow}$ is a function,

$d:R^{pcol\times prow}\times R^{pcol\times prow}\rightarrow R,$ If for all $x,y,z\in R^{pcol\times prow}$ the following axioms are satisfied;

$$N1: d(x,y)\geq 0; d(x,y)=0\Leftrightarrow x=y(\text{positivity}) \qquad (7)$$

$$N2: d(x,y)=d(y,x)(symmetry) \qquad (8)$$

$$N3: d(x,z)\leq d(x,y)+d(y,z)(subadditivity) \qquad (9)$$

It suffices to proof these axioms to establish that Eq.6 is a distance function.

Proof: We note that N1 and N2 can easily be verified and proceed to prove N3. First, we establish *Cauchy Schwartz* inequality for vectors. This states that the inequality (10) holds true for all vectors *x* and *y* of an inner product space.

$$|x.y|\leq|x|.|y| \qquad (10)$$

Assume that, *x=(x₁,x₂,...,xₙ)* and *y=(y₁,y₂,...,yₙ)* and also recall, that the *dot product* of *x* and *y* is given by $x.y=x_1.y_1+x_2.y_2+\ldots+x_n.y_n$. Further, $|x|=\sqrt{x.x}$ and the distance between *x* and *y* in a 1-dimensional space is simply $d(x,y)=|x-y|$.

Now,

$$|x-ay|^2=(x-ay)(x-ay)$$

$$=|x|^2-2a(x.y)+a^2|y|^2 \qquad (11)$$

Using discriminant of Eq.11, we've

$$(-2x.y)^2-4|x|^2|y|^2\geq 0$$

$$\Rightarrow 4(x.y)^2\leq 4|x|^2|y|^2$$

$$|x.y|\leq|x|.|y|$$

Which yields Eq.10 by dividing both si $N3.$ with 4 and taking the square roots. We now use this to proof

For N3,

$$|a+b|^2=(a+b)(a+b)\leq|a|^2+2|a.b|+|b|^2$$

(using Eq.10)

$$\leq|a|^2+2|a||b|+|b|^2=(|a|+|b|)^2$$
$$\Rightarrow|a+b|^2\leq(|a|+|b|)^2 \ and \ |a+b|\leq|a|+|b|$$

R_k^S For any three points/vectors

$$dist_{(s,t)}=|s-t|=|s-r+r-t|\leq|s-r|+|r-t|$$

$$= dist_{(s,r)} + dist_{(r,t)} \quad \text{Q.E.D.}$$

KNN distance hit table: Predicting a test data entails generating a distance 'Hit' Table (1) using Eq.6 and the principle of KNN. Since very data-set is a collection of peaks (vectors), we extended the KNN algorithm to accommodate this. In each interaction, we used Eq.6 to determine the distance between all possible pairs of vectors from the test data-set and train data-set. Then, for each row vector in a test data we noted the stage label of the train data-set vector nearest to it by comparing the k-minimum distance values between the row vectors of train and test data-sets using the distance hit (table 1). Below is an example of a typical hit (table 1).

The column titles *CON*, *MCI* and *tAD* holds counts of the number of rows of stage label's that has k-minimum distance values with respect to a *TEST* data. At the end, a test data is classified into the stage with the highest number of k-minimum hits, e.g., *TEST1* (Table 2) is classified to be *MCI* while *TEST2* is *tAD* based on majority vote.

Using the Jackknife re-sampling technique, each disease stage produced 20 test data-sets. Consequently, weper formed sixty KNN classification iterations with k = 1 and another set of sixty iterations with k = 5. The confusion matrix below is the classification performance obtained with k = 1.

In detail, KNN at k = 1 correctly classified 65% instances of CON data points and correctly classified 40% and 50% instances of MCI and tAD data points respectively, with 10% of tAD elements not conclusively classified.

On the other hand, for the same test samples KNN correctly classified 85% of control(CON) samples, 50% of MCI test samples and 0% of tAD samples with k = 5. Overall, 52% and 45% instances were correctly classified using KNN at k = 1 and k = 5 respectively (Table 3).

RESULTS AND DISCUSSION

The goal of SELDI-TOF discovery proteomics is to quantify and interpret changes in features as to their abundances identified a priori/de novo in the SELDI spectra of analyzed samples by further investigating obtained SELDI Spectra data, inconclusive classifications occurs if an iteration produces equal hit scores for two stages; e.g. $tAD^{\#MCI}$ means an iterations that produced equal hit values for MCI and tAD, for a suspected tAD test data (t) (Table 4).

Table 1: Hit Table.

Hit Table			
	CON	MCI	tAD
TEST1	44	**75**	60
TEST2	49	57	**73**

Table 2: Test 1.

k=1			
	CON	MCI	tAD
CON	**13**	3	4
MCI	5	**8**	7
tAD	3	5	**10**$^{\#MCI}$

In this paper, we adopted the principle of KNN and introduced a 2-scale distance function to build a KNN classifier for Alzheimer's disease stages based on the molecular mass and TOF-Intensity of ions contained in SELDI Saliva Spectra data. This study forms a basis and provides a pathway into studies on early and reliable diagnoses of AD and Dementia disease in general.

The data structure was the first problem we had to overcome. The decomposition of the feature matrices to a collection of feature vectors, sequel to a systematic feature selection enabled us to solve the problem in a2-dimensional space.

This work pinpoints inherent pattern in the saliva SELDI data. The results of 5-NN algorithm on tAD test data points, clearly indicates a characteristic 'elusiveness' possessed by the data, which can be explained by the lack of cognition suffered by Alzheimer's disease (Dementia) patients in general. On the other hand, it further proves the reliability of SELDI process and

Table 3: Test 2.

k=5			
	CON	MCI	tAD
CON	**17**	3	0
MCI	10	**10**	0
tAD	16	4	**0**

Table 4: The default diagnosis is diagnosis done by flipping a coin. An unbiased expected output is presented by the confusion matrix below;

Default Diagnoses		
Predisposed	NO	YES
NO	**10**	10
YES	**20**	**20**
p(type 2 error)=0.33		
50% Accuracy		

Table 5: Through default diagnosis the probability of committing type II error is 0.33

Predisposed (*k=1*) Diagnoses		
Predisposed	NO	YES
NO	**13**	7
YES	**8**	**32**
p(type 2 error)=0.13		
75% Accuracy		

reproducibility of mass spectra results as studied by Keith et al., [16].

If combined with clinical records and coupled with clinical verifications, the result of this study forms a basis for discriminating and diagnosing Alzheimer's disease. It can also serve as a tool to monitor AD patients conditions since the disease severity status can easily be determined with the number of 'hit points' in the KNN distance table, knowing that, the distance measure between two vectors remains the same except if there is a change in the geographical locations of one or both of them.

Using Saliva SELDI data was also a plus owing to how easily saliva samples can be obtained. The presence of several molecular mass values but with different intensity values made this KNN approach possible, in that, we were able to geometrically mark the intensities of similar mass values in space and used their geometric location for discrimination.

There are classification scenarios that need to be further explained clinically in terms of false negative predictions/ classifications. For instance, if the hit point scores for two stages (e.g. MCI and tAD) are the same, what should be the result of such classification? To a layman, this indicates a YES to the question of having the disease.

By virtue of the *delete-1* Jackknife procedure, every data-point in the data-set was in turn tested. Thus, we exhaustively evaluated the model's performance. This was handy since the data is small in size and works on saliva SELDI data-set is not available in literature.

A further interpretation of our result is two way classifications; *viz-a-viz* predisposed and indisposed persons. Consider k=1 confusion matrix again, notice that members of MCI stage has the tendency to exhibits 50% characteristics of both CON and tAD as evident in the proportion of miss classified instances of MCI. Owing to this, let's regroup the sampled population; CON as non-predisposed and MCI and tAD as predisposed and compare the tendency of committing *type II error* (Table 5) based on *k* = 1 classification against result obtained via default diagnosis of the same population size.

CONCLUSION

It is worthy to ask if adding additional features into the distance function will improve the result of this work. Similarly, will it improve the obtained result if only ions of a particular peak type are used or if ions are categorized and used based on their molecular weight or signal to noise ratio?

The model described here was done with SELDI saliva data set generated with CM10 (cation exchange surface) chemistry at low (1800 nJ) laser energy bombardment condition, as another possible area of future work, one could extend this work to other SELDI data generated under other energy conditions and/or chemistry. A sensitivity analysis of saliva SELDI data with regards to the best time of the day saliva samples can be obtained from donors for SELDI examination is also a possible future work.

Having transformed the matrix data points into a collection of row vectors, building and determining the performance of other models with other learning algorithms including applying the statistical distribution of mass and corresponding TOF Intensity values of ion molecules as expressed across stages can be looked at as a future work.

In conclusion, while studies aimed towards personalized medicine are currently on going, the focus on closing the gap between bio-marker identification and the diagnosis of incurable diseases (e.g Dementia) should not be lost.

REFERENCES

1. Issaq H, Veenstra T, Conrads T, Felscow D. The SELDI-TOF MS Approach to Proteomics: Protein Profilingand Biomarker Identification. Biochem Biophys Res Commun. 2002; 292: 587-597.
2. Raghava GPS. Bioinformatics and drug discovery. 2015.
3. Jonathan M Street, James W Dear. The Application of Mass Spectrometry Based Protein BiomakerDiscovery to Theragnostics. Br J Clin Pharmacol. 2010; 69: 367-378.
4. Alzheimer's Association.
5. Jae Won Lee, Jung Bok Lee, Mira Park, SeuckHeun Song. An extensive comparison of recent classificationtools applied to microarray data. Comput Sta Data Anal. 2005; 48: 869-885.
6. Dudoit S, Fridlyand J, Speed P. Comparison of discrimination methods for classification of tumors usinggene expression data. JASA. 2002; 97: 77-87.
7. José Crossa, Jorge Franco. Statistical Methods for Classifying genotypes. Euphytica. 2004; 137: 19-87.
8. Li L, Weinberg CR, Thomas A. Darden and Lee G. Pedersen. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameter of GA/KNN method. Bioinformatics. 2001; 17: 1131-1142.
9. Min-Ling, Zhi-Hua Zhou. ML-KNN: A lazy learning approach to multi-label learning. Pattern Recogn. 2007; 40: 2038-2048.
10. Wangmeng Zuo, David Zhang, Kuanquan Wang. Bidirectional PCA with assembled matrix distance for image recognition. Cybernetics. 2006; 36: 863-872.
11. Holm L, Sander C. Protein Structure Comparison by Alignment of Distance Matrices. J Mol Biol. 1993; 233: 123-138.
12. Kilian Q. Weinberger, Lawrence K. Saul. Distance Metric Learning for Large Margin Nearest Neighbor Classification. JMLR. 2009; 10: 2007-2244.
13. Hao Zhang, Alexander C. Berg, Michael Maire, Jitendra Malik. SVM-KNN: Discriminative Nearest NeighborClassification for Visual Category Recognition. IEEE. 2006; 2: 2126-2136.
14. Avery I. McIntosh. "The Jackknife Estimation Method". 2016.
15. Sandra L. Taylor, Kyoungmi Kim. A Jackknife and Voting Classifier Approach to Feature Selection and Classification. Cancer Inform. 2011; 10: 133-147.
16. Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF protein patterns in serum: Comparing datasets from different experiments. Bioinformatics. 2004; 20: 777-785.

Cite this article

Anyaiwe DEO, Wilson GD, Geddes TJ, Singh GB (2016) K-NN Classification of Mass Spectra Data for Diagnosing Alzheimer's Disease. Comput Sci Eng 1(1): 1004.