Enhancing Classification of Mass Spectrometry Imaging Data with Deep Neural Networks

Spencer A. Thomas*, Yaochu Jin†, Josephine Bunch*‡, Ian S. Gilmore*
*National Centre of Excellence in Mass Spectrometry Imaging (NiCE-MSI)
National Physical Laboratory, Hampton Road, Teddington, TW11 0LW, UK
†Department of Computer Science, University of Surrey, Guildford, GU2 7XH, UK
‡Imperial College London, London, SW7 2AZ, UK
Corresponding Author: spencer.thomas@npl.co.uk

Abstract—**Mass spectrometry imaging (MSI) determines the spatial distribution of thousands of molecules and chemical species simultaneously and has emerged as a powerful suite of tools in pathology, pharmaceuticals, healthcare and material science applications. MSI experiments typically produce between 10^4 and 10^6 pixels, each of which contain a full mass spectrum. The use of MSI in pathological applications may involve classification of tissues based on the spectral information, though to date there have been no systematic studies of classification algorithms, comparison of performance when using the full versus reduced dimensionality of the data, or investigation into multi-class classification problems. Here we evaluate a number of algorithms for classifying regions in a MSI dataset before and after unsupervised non-linear dimensionality reduction with a deep neural network. We evaluate the performance of each algorithm with eight metrics in both the high (1,601) and low (3) dimensional feature space. Our results show that in the high dimensional space, only a Softmax classifier and support vector machine (SVM) with a linear kernel, perform to a satisfactory level, with all other algorithms overfitting the training data and performing poorly on the testing data. We also observe that that multi-class classification performance is drastically improved by the non-linear reduction with the deep neural network, improving scores and reducing variation between classes compared with the original high dimensional data. In the low dimensional space, the best performance observed is from a Decision Tree, though KNN, SVM (Gaussian kernel) and Softmax classifiers also perform well, scoring over 0.93 across all metrics and classes.**

I. Introduction

Classification of high dimensional data is difficult and computationally intensive. The computational challenges are compounded for large datasets. Mass spectrometry imaging (MSI) is a suite of chemical imaging techniques that are able to measure the distribution of a wide range of chemical species (organic and inorganic) and specific locations (pixels) in a sample. Depending on technique and instrumentation, the resolution of spatial and chemical information can vary. For each pixel there is a full mass spectrum which measures the mass-to-charge ratio (*m/z*) of the components removed from the surface of the sample. The range and resolution of this spectrum can also vary with technique and instrumentation. Typically the number of pixels range from 16,000 to over 4,000,000 and the number of spectral channels can ranges between 100s and 1,000,000s of *m/z* intensities. As a consequence, these data are high dimensional but typically sparse and often contain redundant information of the features present, thus data reduction techniques and feature extraction are typically employed. Classification studies with mass spectrometry data typically use a single algorithm for binary classification problems. Multi-class problems are lacking as are investigation into the performance of various classification algorithms for mass spectrometry data.

MSI data has proven extremely useful in a wide range of studies inducing material characterisation [1]–[3], clinical and pathological [4]–[6], pharmaceutical [7]–[9], environmental [10], and forensic sciences [11], [12]. However, classification of MSI data based on the full spectrum is both difficult and unnecessary. As the data are typically very sparse, much of the data can be uninformative and the problem may be simplified. Noise, experimental fluctuations and pixel-to-pixel variations can lead to stark differences in neighbouring pixels. This is typically observed as changes in the signal intensity, loss of a signal, or increased noise. Chemical differences in the sample can also lead to subtle differences in the spectra, such as a change in the ratio between signals, or the additional presence of low intensity signals across the mass range.

Studying latent or reduced variables has been shown to be an effective way of analysing data from high dimensional or complex system [13]. A number of dimensionality reduction methods currently exist and are used in MSI data analysis. Methods such as principal component analysis (PCA) and non-negative factorisation (NMF) are fast but assume linearities in the data in standard implementations which may not hold true in MSI data [14]. Similarly, clustering based methods are routinely applied to MSI data. Although non-linear kernels can be used, issues such as the number of factors required to reconstruct or segment the data remain. Over the last few years t-distributed stochastic neighbour embedding (t-SNE) has emerged as a powerful non-linear dimensionality reduction technique [15], [16] and has been used in MSI imaging for features extraction [17]–[20]. However, even with approximations that reduce the complexity [21], t-SNE is still prohibitive for large MSI datasets. Many of these cases require construction and evaluation of a covariance or distance matrix, iterative calculations, or lack a mapping between the high and low dimensional spaces. The memory requirements set a hard limit for the size of a dataset that will depend on the

978-1-5386-2726-6/17/$31.00 ©2017 IEEE

RAM available for the computation, while the lack of mapping prevents compatible application across datasets.

Recently, the use of an autoencoder, a symmetric neural network, has been applied to MSI data for dimensionality reduction [14]. This enables non-linear dimensionality reduction through the use of a sigmoid or tanh activation function with a fixed mapping between the high and low dimensional space. This permits the unsupervised training of the network based on a training set from the data which can be applied to the full dataset once trained. Moreover, by only requiring a subset of the data for training, the full dataset is not required to be held in memory during a given training epoch, providing a memory-efficient non-linear dimensionality reduction technique. The ability to perform training on a subs set of the data permits the analysis of much large MSI datasets than current methods.

In this paper we briefly introduce MSI data in Section II. We then overview the classification algorithms, performance metrics and dimensionality reduction used in this work in Section III. Finally we present the results and conclusions of our investigation in Section IV and Section V respectively.

II. Mass Spectrometry Imaging Data

Mass spectrometry (MS) is a collection of techniques that remove material from the surface of a sample typically through ionisation or desorption processes which are passed to a mass analyser in order to determine their mass to charge ratio (*m/z*). This forms a spectra with peaks that correspond to the ionised material. These peaks can be used to identify the chemical composition of a material through the detection of intact molecular species, or fragments of them, from typically sparse and noisy spectra. Peaks in the spectrum indicate possible molecular components of the, potentially unknown, material. Although efforts to automatically combine analysis with reference databases using data mining techniques, in many cases peaks are labelled manually.

In mass spectrometry imaging (MSI), a mass spectrum is acquired at regular spatial locations of a sample, typically in two dimensions. This builds up an image of the molecular distribution of a sample, as each pixels is a mass spectrum. The ability to detect the spatial distribution of a wide range of unlabelled chemical species has led to MSI becoming an important tool in many application areas such as studies of drug effects in tissue, histopatholgy and disease classification [4]–[9]. This increases the need for data-driven analysis as correlation, causation, or variations may be unknown.

MSI can provide both qualitative and sometimes quantitative differences in the chemical composition of a sample under different conditions. The ability to capture material composition and distribution has made MSI techniques extremely useful in analysing biological samples as they can identify drugs [22], [23], lipids [24]–[26], peptides [27], proteins [28] and metabolites [22], [23] from many different tissues. Here we use a secondary ion mass spectrometry (SIMS) MSI dataset which can provide high lateral resolution images of molecular distributions [29] even down to the cellular level [30], [31]. Specifically the data in this work is a SIMS image of a mouse lung consisting of over 98,000 pixels and 1,601 *m/z* channels after applying a noise threshold to the spectra. We evaluate the ability of various algorithms detailed below at distinguishing five different regions in the data based on the similarity of chemical composition measured by MS.

III. Classification

A number of algorithms exist for classification problems with several reviews available in the literature [32]–[34]. Some have been applied to mass spectrometry data, typically in the context of predicting disease tissue. However, to date there has not been a systematic investigation in to; the suitability of the classification algorithms to MSI data; a study of multi-class classification; the performance difference in the full and reduced dimensional space; or the use of autoencoders for the non-linear mapping.

A. Algorithms

1) K-Nearest Neighbours (KNN): A label is assigned to a test cases based on the most frequent label of its closest k-neighbours as defined by a distance metric. Here we select Euclidean distance and k=1 meaning a test case is assigned the same label as the training instance it is closest to.

2) Linear Discriminant Analysis (LDA): The features that describe the data are transformed to a linear combination of the features that maximises the discrimination of the classes in the data. The linear transform is defined as

$$\hat{\mathbf{x}} = \mathbf{x} \times \mathbf{W} \ , \qquad (1)$$

where $\mathbf{W}$ is the projection matrix. The optimal projection will maximise between class scatter matrix, S_b, while minimizing the within class scatter matrix, S_w. These are defined as

$$S_w = \sum_{j=1}^{C} \sum_{i=1}^{N_j} (x_i^j - \mu_j)(x_i^j - \mu_j)^T \ , \qquad (2)$$

$$S_b = \sum_{j=1}^{C} (\mu_j - \mu)(\mu_j - \mu)^T \ , \qquad (3)$$

where C is the number of classes, μ is the mean of all classes, and x_i^j, μ_j and N_j are the ith, mean and total number of instances in class j respectively. If S_w is non singular, the optimal projection can be obtained when $\mathbf{W}$ corresponds to the eigenvectors of $S_w^{-1} S_b$. Here we select eigenvectors that correspond to the 10 largest eigenvalues of $S_w^{-1} S_b$.

3) Support Vector Machine with Linear Kernel (SVM-L): Support vector machines construct a hyperplane that provides the greatest margins between instances belonging to different classes in the training data. The training examples on these margins are referred to as support vectors and the hyperplane is defined as

$$\hat{\mathbf{x}} = \mathbf{W} \cdot \mathbf{x} - b \ , \qquad (4)$$

where $\mathbf{W}$ is a normal vector to the hyperplane. The support vectors are obtained by solving Eq. (4) for $\hat{\mathbf{x}} = \pm 1$. The classifier is trained by maximising the separation between the support vectors under the constraint that as few instances lie between the margins as possible and the distance between them and the margins is small. This so called *soft margin* is achieved by minimising the loss function

$$L(\hat{x}) = \sum_i \max\left(0, 1 - \hat{\mathbf{x}}_i(\mathbf{W}^T\mathbf{x}_i) - b)\right) \ , \tag{5}$$

during training of the SVM classifier.

4) SVM with Gaussian Kernel (SVM-G): For data that are not linearly separable, a non-linear classifier is required. For SVM this can be achieved by replacing the dot products in the original linear algorithm with a non-linear function known as a kernal trick. This enables a linear separation following a non-linear transformation, or projection to a high dimensionality where the data are linearly separable. Here we select a Gaussian function defined, for $\gamma \geq 0$, as

$$\Psi(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2} \ . \tag{6}$$

The separation task is now defined as

$$\hat{\mathbf{x}} = \sum_i \alpha_i \Psi(\mathbf{x}_i, \mathbf{x}) - b \ , \tag{7}$$

where α_i is a parameter, $\mathbf{x}_i$ and $\mathbf{x}$ are a feature and support vectors respectively.

5) Naive Bayes with Gaussian fit: We construct a Bayesian probabilistic model based on the training examples in the form

$$p(C_j|\mathbf{x}) = \frac{p(C_j)\ p(\mathbf{x}|C_j)}{p(\mathbf{x})} \ . \tag{8}$$

This model makes the assumption that the different classes, and therefore features, are independent of each other, referred to as a *Naive* assumption. The likelihood function $p(\mathbf{x}|C_j)$ is modelled with a Gaussian distribution based on the mean and variance for each of the classes.

6) Naive Bayes with Kernel Smoothing fit: Here we use a Naive Bayes classifier as before, but using a kernel smoothing density estimator in place of a Gaussian distribution to model the likelihood function. We select a uniform kernel defined as $f(\mathbf{x}) = 0.5\ l\{|\mathbf{x}|\} \leq 1$, where $l\{u\}$ is the indicator function denoting the class membership of u.

7) Softmax Function: The softmax function compresses the elements of a real valued input vector $\mathbf{x}$ such that $\sum_i \hat{\mathbf{x}}_i = 1$. For an test instance the output of the softmax function is the probability that it belongs to each class and the labelled assigned is based on the highest probability returned. The softmax classifier for the C classes is defined as

$$P(\hat{\mathbf{x}}_i = c|x; \mathbf{W}) = \frac{e^{\mathbf{x_i^T W_c}}}{\sum_j^C e^{\mathbf{x_i^T W_j}}} \ , \tag{9}$$

where $\mathbf{W}$ is a weight vector of C dimensions, resulting in a distinct linear operator for each class. The classifier is optimised by updating $\mathbf{W}$ such that it minimises the cross entropy, E, of the function for N training examples and C classes.

$$E = \frac{1}{N}\sum_i^N \sum_j^C y_i \ln \hat{x}_{ij} + (1 - y_i)\ln(1 - \hat{x}_{ij}) \ , \tag{10}$$

where y_i is the true label of the ith training example.

8) Decision Trees (DT): A decision tree is constructed such that it splits the (training) data in accordance with their labels. The tree structure is modified to minimize the Gini impurity, which for N training examples is defined as

$$I_G(p) = \sum_j^C p_j(1 - p_j) \ , \tag{11}$$

where p_j is the probability of instances labelled with class j.

B. Performance Metrics

There are a number of metrics to assess the performance of classification algorithms. To enhance the generality of our study, we consider multiple metrics for each algorithm across a multi-class classification problem. For the number of instances that are classified as true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN), by a given algorithm, we can define the following performance metrics.

$$\textit{Sensitivity} = \frac{TP}{TP + FN} \tag{12}$$

$$\textit{Precision} = \frac{TP}{TP + FP} \tag{13}$$

$$\textit{Specificity} = \frac{TN}{TN + FP} \tag{14}$$

$$\textit{Accuracy} = \frac{TP + TN}{TP + FN + TN + FN} \tag{15}$$

$$\textit{F-Measure} = \frac{2TP}{2TP + FP + FN} \tag{16}$$

$$\textit{Informedness} = \textit{Sensitivity} + \textit{Specificity} - 1 \tag{17}$$

$$\textit{Markedness} = \textit{Precision} + \frac{TN}{TN + FN} - 1 \tag{18}$$

$$\textit{Matthews Correlation Coefficient (MCC)} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{19}$$

Note that all metrics only have values assigned if they have non-zero denominators. If a metric is computed as not a number (NaN) no score is given, with a value of zero returned for that metric.

C. Dimensionality Reduction

Due to the high dimensionality of MSI data we investigate the use of dimensionality reduction prior to classification in a MSI data context. Previous studies have manually selected features to classify (non-imaging) mass spectrometry data [35], [36], which is useful if one knows all the informative features in the data. However, in the case where important features or peaks in the data are not known, which is typical for MSI experiments, these methods are not appropriate. The ability to measure 1000s of unknown molecules simultaneously is a significant strength of these techniques and thus prohibits supervised learning techniques due to the lack of *a priori* information in the data. Recently Nguyen et al [37] used wavelets coupled with genetic algorithms to perform dimensionality reduction prior to a binary classification problem. Other studies have focused on the selection of appropriate features from high dimensional data so that the inclusion of uninformative features does not results in classifier degradation [38].

Methods such as PCA, clustering and factorisation based methods, which are typically linear, require the selection of a number of features in the data for reconstruction or features selection. This requires knowledge of the data or potentially iterative applications to achieve satisfactory results. Non-linear methods such as t-SNE and neural networks can generalise the features in the data and map all of the high dimensional data in to a low dimensional space, typically two or three features, without knowledge of the dataset. Typically t-SNE is able to provide a low dimensional map that outperforms neural networks (in terms of decreasing inter-group, and increasing intra-group distances) [15]. The use of t-SNE for classification of tumour sub-populations has been recently investigated by Abdelmoula et al. [20] with impressive results. However, the computational complexity of this algorithm, even with the Barns-Hut approximation [21], prohibit this algorithm from being applied to large individual MSI images or tiled combined datasets. Moreover, the mapping obtained from t-SNE between the high and low dimensional spaces is unknown and can not be transferred to other datasets or reproduced unlike for the neural network [14]. The fixed mapping of this transfer in neural networks permits investigation that relate the encoding and classification back to the original data.

The untargeted measurements in the MSI experiment provide a wealth of spatial and chemical information about the sample though lacks any annotations of chemical features therefore prohibits any supervised learning tool. Therefore we use a symmetric neural network, an autoencoder, which is able to perform unsupervised non-linear dimensionality reduction. This method has been shown to be successful in MSI data for reduction [14]. Autoencoders have the additional advantage of the ability to use a subset of the data for training enabling application to larger imaging datasets where other methods are prohibited.

We use a deep architecture of sparse stacked autoencoders trained in a greedy layer wise fashion. The high dimensionality of MS data, which can contain 10^6 mass channels, can potentially lead to a prohibitively large number of parameters to train in an end-to-end fashion. Moreover, layer wise training has been shown to be successful in dimensionality reduction of MSI data [14]. This permits unsupervised learning by training the network to first encode, then decode an input signal, using gradient descent to minimize the error between the input and decoded signal. The input vector $\mathbf{x}$ represents the mass spectrum for a given pixel and is encoded to the latent variable space $\mathbf{z}$ using a sigmoid activation function, σ.

$$\mathbf{z} = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b}) \ . \tag{20}$$

Here $\mathbf{W}$ and $\mathbf{b}$ are the network weight matrix bias vector respectively. The latent variables are then decoded using a similar operation

$$\mathbf{x}' = \sigma'(\mathbf{W}'\mathbf{z} + \mathbf{b}') \ . \tag{21}$$

The network is trained using gradient descent to minimize the reconstruction error $\epsilon = ||\mathbf{x} - \mathbf{x}'||^2$. We employ a standard L_2 norm weight regularisation for w_{ij}, the elements of $\mathbf{W}$. This is calculated for L hidden nodes, n training examples and k variables in the training data, here the spectral channels,

$$\Omega_w = \frac{1}{2} \sum_{l}^{L} \sum_{j}^{n} \sum_{i}^{k} \left(w_{ji}^{(l)} \right)^2 \ , \tag{22}$$

and include sparsity regularisation using the Kullback-Leibler (KL) divergence between the ρ and $\hat{\rho}$, the desired and observed sparsity respectively.

$$\Omega_s = \sum_{i=1}^{D'} \rho \log \left(\frac{\rho}{\hat{\rho}_i} \right) + (1 - \rho) \log \left(\frac{1 - \rho}{1 - \hat{\rho}_i} \right) \ , \tag{23}$$

where D' is the number of reduced dimensions. The total cost function used to train the regularised sparse autoencoder is given as

$$E = \frac{1}{n} \sum_{j=1}^{n} \epsilon_j + \lambda \Omega_w + \beta \Omega_s \ . \tag{24}$$

The deep network architecture used in this work consists of an input layer for the 1,601 *m/z* intensity channels, hidden layers of 100, 20, 10 and 3 nodes respectively in a symmetric layout, and finally an output layer of the reconstructed 1,601 *m/z* values. We used the encoded data in the lowest dimensionality, here three, to visualise the mouse lung dataset (Fig. 1) and assign the labels detailed in the next section.

D. Labelling

Using the deep autoencoder described above, we reduce the dimensionality of the MSI data from 1,601 *m/z* channels to three deep features. We can visualise this as a red blue green (RGB) images using each of the deep features as a colour channel intensity. Areas in Fig. 1 with the same colour correspond to pixels with similar spectra, that is areas in the sample with the same chemical composition. Visualising the data in this way enable these regions to be distinguished from each other via different colours in the RGB space. We can

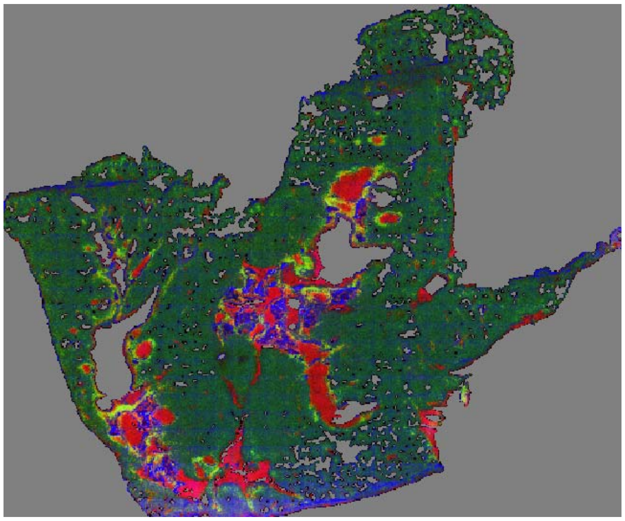


Fig. 1. Visualisation of the 1,601 *m/z* channel MSI dataset as an RGB image after unsupervised non-linear dimensionality reduction using a deep autoencoder reduces the 1,601 *m/z* channels to just three values.

TABLE I
CLASS INSTANCES IN TRAINING AND TESTING DATA

Dataset	Class 1	Class 2	Class 3	Class 4	Class 5
Training	855	852	956	210	7742
Testing	7760	7586	8610	1918	69655
Total	8615	8438	9566	2128	77397
Colour (Fig. 1)	Red	Green	Blue	Black	Dark Green

therefore assign labels to the pixels in the MSI dataset based on their colour as it represents similar chemical composition. We assign the class of a pixel in the data based on the colour with the number of instances listed in Table I for each of the colours in Fig. 1. These labels allow us to construct a multi-class classification problem in both the high (1,601) and low (3) dimensional spaces using the algorithms listed in Section III-A. Note that the encoded data represents the deep features present in the high dimensional space, but has not changed the features contained in each pixel. That is the low dimensional representation has reduced the redundancies in the data present in the high dimensional space.

E. Training and Testing Data

The original dataset of 106,144 pixels is divided into training and testing data, where the training data contains approximately 10% of the data from each class. This way we are able to test the algorithm's ability to generalise based on limited data, retain the class imbalance and not to bias the classifiers during training. After training we then apply the classifiers to the unseen testing data for validation. We apply the metrics to both the training and testing data to compare the algorithm's performance across each of the classes. A summary of the breakdown of the training and testing data sizes for each class is given in Table I.

IV. RESULTS

The class averaged metric scores for the training data are given in Fig. 2 for all classifiers in this work with error bars representing the standard deviation of the metric's score over the five classes. The scores for classifiers in the lower dimensional space increase due the removal of redundancies present in the higher dimensional data. This is true for all algorithms, with SVM-Linear being an exception. The performance of this algorithm is hindered by its inability to correctly classify any instance in the test set from the smallest class (class 4) which lowers the average performance and results in a very large standard deviation. For this algorithm we also include the class averaged scores excluding class 4 indicated by the crosses in Fig. 2(c), which shows a similar trend to the other algorithms when comparing high and low dimensional data.

When considering the high dimensional data, only the SVM-Linear and Softmax classifiers perform well, scoring over 0.64 for all metrics and classes. SVM with a linear kernal drastically outperforms the Gaussian kernal version which yields highly variable and low scores in all metrics other than accuracy. For the low dimensional data, the Decision Tree, KNN, SVM-Gaussian and Softmax classifiers perform best scoring over 0.93 in all class for each metric. The Decision Tree based algorithm produces the best performance scoring over 0.99 in all cases.

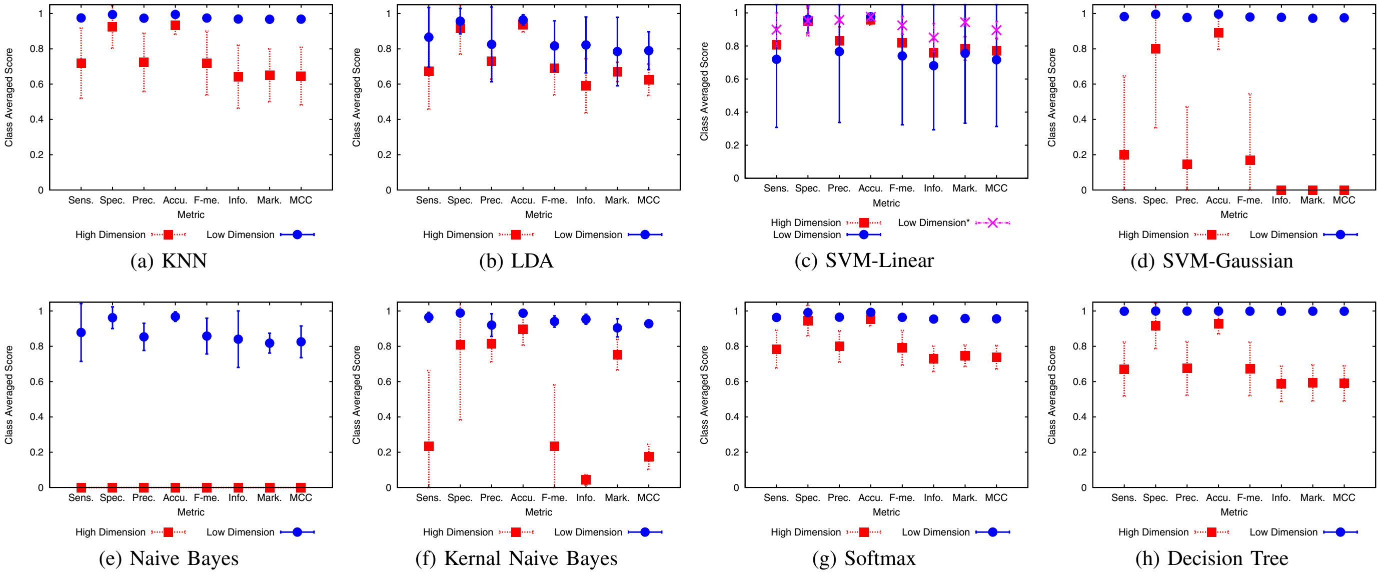
The between class variability has been drastically reduced for the low dimensional data for all algorithms, apart from SVM-Linear as noted above unless class 4 is removed. This indicates that the use of the non-linear reduction has circumvented any effects from within and between class variance on algorithm performance.

The low dimensional data also permits the usage of the Naive Bayesian classifier which failed to build a prediction model of the high dimensional data. This resulted from an inability to fit a Gaussian distribution to the classes implying that the data are not normally distributed in the original high dimensional space. When using a kernal smoothing fit for the predictive model in a Naive Bayes classifier, the performance improves but is still very poor compared to other algorithms. This classifier is more computationally intensive than the other algorithms due to the smoothing fit. The difficulties in fitting a Gaussian distribution to the data also explains why the SVM-Gaussian algorithm performs poorly for the high dimensional data. The SVM-Gaussian and both Bayesian classifiers show significant enhancement of performance scores and even enabling for some cases. The KNN, LDA, Softmax and Decision Tree algorithms also show a large enhancement in performance across all metrics.

As an additional level of analysis we also look at the performance ratio for the class averaged metrics defined as

$$\phi_i = \frac{\text{testing score}_i}{\text{training score}_i} \ , \tag{25}$$

where i denotes a given metric. The ϕ value is an indicator of the level of overfitting in the classifier, where ϕ=1 denotes the same performance of an algorithm in both the training



(a) KNN

(b) LDA

(c) SVM-Linear

(d) SVM-Gaussian

(e) Naive Bayes

(f) Kernal Naive Bayes

(g) Softmax

(h) Decision Tree

Fig. 2. Evaluation of classifier performance on the MSI testing data in the full (high) dimensional space in dashed red and encoded (low) dimensional space in solid blue. Values here reflect the testing scores for each metric averaged over the five regions in Fig, 1. The error bars are the standard deviation of metric scores over the five classes. The additional points in (c) are the mean and standard deviation over classes 1,2,3 and 5 as the classifier failed on all instances of class 4, the smallest class, in both training and testing sets.

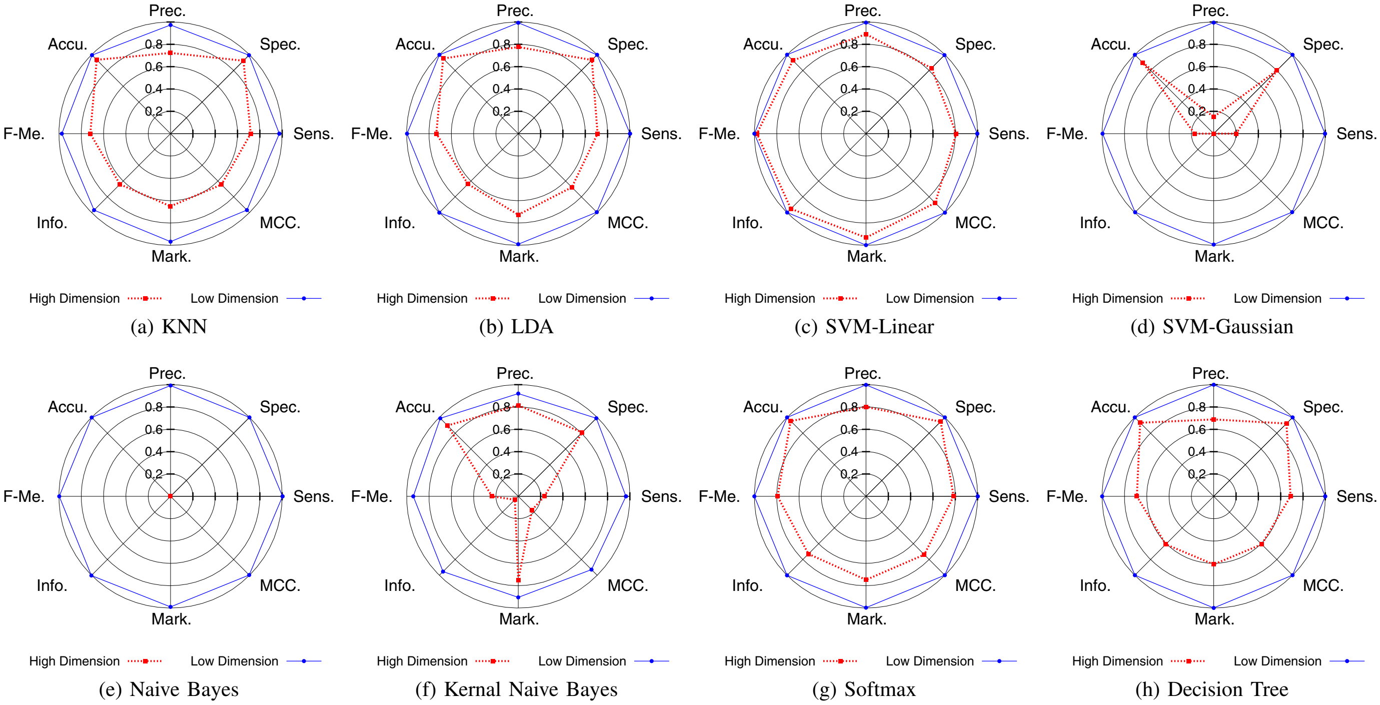
and testing data sets. As we use ≈90% of the data in each class for testing, $\phi \approx 1$ is a good indicator that the algorithm has learnt the general patterns in the data. The ϕ values of all metric for each algorithm is illustrated Fig. 3. Here we can see that all algorithms are overfitting the high dimensional training data to some degree, where $\phi < 0.8$ for at least one metric. In the low dimensional space, however, overfitting is reduced in all algorithms with $\phi \geq 0.9$, with the Decision Tree, Softmax and both SVM algorithms achieving $\phi > 0.99$ in all cases (classes and metrics). Comparing the results from Fig. 2(c) and Fig. 3(c) we can see that although the linear SVM algorithm can not correctly label instances from the smallest class in the test set for the low dimensional data, this is also true for the training data as indicated by the ϕ values in Fig. 3 providing more insight in to the algorithm.

These results verify our multi-metric evaluation of this problem as it is well known that some metrics do not provide a complete picture of a classifier's performance. As demonstrated in our results in Fig. 3, ϕ values for Accuracy and Specificity would indicate the algorithms are generalising the high dimensional data well, and algorithms typically score highly in these metrics (see Fig. 2). However, ϕ values for almost all other metrics are below 0.8 indicating overfitting. The results in Fig. 3 support our conclusions from Fig. 2 that only the SVM-linear and softmax classifiers perform at a satisfactory level when using the high dimensional data. Figure 3 also illustrates the failings of the SVM-Gaussian and Naive Bayes (Gaussian and Kernal) Classifiers.

V. Conclusions

In this work we have compared a range of algorithms for the multi-class classification of a MSI dataset. We evaluate the performance using a number of well established metrics in the field of classification. Our results clearly show that the multi-class classification of MSI data is enhanced, even enabled, when performing non-linear dimensionality reduction using a deep autoencoder prior to training a classification algorithm. In the full dimensional space there is sever overfitting in a number of well established algorithms, and generally poor performance based on the eight metrics used in this work. Our results indicate that only the Softmax and SVM-L algorithms are suitable for classification of MSI when considering the original high dimensional data as both algorithms scored over 0.6 in all cases (metrics and classes). However, these results strongly suggest that (non-linear) dimensionality reduction should be performed prior to classification to improve performance and reduce overfitting.

The dimensionality reduction step increases the scores from each of the metrics across all classes, even in the presence of a significant imbalance with the smallest and largest classes accounting for 2% and 73% of the total data respectively. Moreover, for a given algorithm, the variation between the five classes for a given metric was drastically reduced. We also observed that four of these algorithms, Decision Tree, KNN, SVM-G and Softmax perform well in all cases (classes and metrics) consistently scoring over 0.93, with the Decision Tree observed as the overall best in the experiments in this work. The drastic improvement in classification performance in the low dimensional space is due to the ability of a deep autoencoder to reduce redundancies in the data and capture all the features present in the high dimensional data. The fixed network provides a mapping between the low and high dimensions enabling the spectra to be directly related to the classification performance. Furthermore, this fixed mapping



(a) KNN (b) LDA (c) SVM-Linear (d) SVM-Gaussian

(e) Naive Bayes (f) Kernal Naive Bayes (g) Softmax (h) Decision Tree

Fig. 3. Evaluation of classifier performance using ϕ defined in Eq. 25 for MSI data in the full (high) dimensional space in dashed red and encoded (low) dimensional space in solid blue. Values here reflect the class averaged ϕ value for each metric. Visualising the ratio of testing and training metrics shows clearly that all classifiers are either overfitting or failing in the high dimensional space, both of these are circumvented when using the encoded low dimensional data.

enables the transfer of the low dimensional mapping learnt by the network to other datasets yielding a comparable low dimensional representation between different instances of the data. That is, we can train the autoencoder on a subset of the data and apply the learned encoding to the remaining data as demonstrated here.

We have presented a systematic comparison on several algorithm's performance in multi-class classification problems. Our results indicate that classifiers trained on the low dimensional data are less prove to overfitting. Only the SVM-Linear algorithm in the low dimensional space appears to struggle with class imbalance with all other algorithms performing well over all classes. We observe the best performance, both in terms of highest score and smallest between class variation, from the Decision Tree, KNN, SVM-Gaussian and Softmax classifiers.

Acknowledgment

This work forms part of the Strategic Capability Programme (projects "3D OrbiSIMS" and "AIMS-HIGHER") of the National Measurement System of the UK Department of Business, Energy and Industrial Strategy. Thanks to Melissa K. Passarelli (NPL) for providing the dataset used in this work.

References

[1] M. J. Bailey, R. Bradshaw, S. Francese, T. L. Salter, C. Costa, M. Ismail, R. P. Webb, I. Bosman, K. Wolff, and M. de Puit, "Rapid detection of cocaine, benzoylecgonine and methylecgonine in fingerprints using surface mass spectrometry," *Analyst*, vol. 140, pp. 6254–6259, 2015.

[2] R. Havelund, A. Licciardello, J. Bailey, N. Tuccitto, D. Sapuppo, I. S. Gilmore, J. S. Sharp, J. L. S. Lee, T. Mouhib, and A. Delcorte, "Improving secondary ion mass spectrometry c60n+ sputter depth profiling of challenging polymers with nitric oxide gas dosing," *Analytical Chemistry*, vol. 85, no. 10, pp. 5064–5070, 2013.

[3] C. Fleischmann, T. Conard, R. Havelund, A. Franquet, C. Poleunis, E. Voroshazi, A. Delcorte, and W. Vandervorst, "Fundamental aspects of arn+ sims profiling of common organic semiconductors," *Surface and Interface Analysis*, vol. 46, no. S1, pp. 54–57, 2014, sIA-13-0492.R1.

[4] N. Abbassi-Ghadi, E. A. Jones, K. A. Veselkov, J. Huang, S. Kumar, N. Strittmatter, O. Golf, H. Kudo, R. D. Goldin, G. B. Hanna, and Z. Takats, "Repeatability and reproducibility of desorption electrospray ionization-mass spectrometry (desi-ms) for the imaging analysis of human cancer tissue: a gateway for clinical applications," *Anal. Methods*, vol. 7, pp. 71–80, 2015.

[5] J. Seuma, J. Bunch, A. Cox, C. McLeod, J. Bell, and C. Murray, "Combination of immunohistochemistry and laser ablation icp mass spectrometry for imaging of cancer biomarkers," *PROTEOMICS*, vol. 8, no. 18, pp. 3775–3784, 2008.

[6] R. L. Edwards, A. J. Creese, M. Baumert, P. Griffiths, J. Bunch, and H. J. Cooper, "Hemoglobin variant analysis via direct surface sampling of dried blood spots coupled with high-resolution mass spectrometry," *Analytical Chemistry*, vol. 83, no. 6, pp. 2265–2270, 2011.

[7] J. Bunch, M. R. Clench, and D. S. Richards, "Determination of pharmaceutical compounds in skin by imaging matrix-assisted laser desorption/ionisation mass spectrometry," *Rapid Communications in Mass Spectrometry*, vol. 18, no. 24, pp. 3051–3060, 2004.

[8] K.-C. Schäfer, J. Balog, T. Szaniszl, D. Szalay, G. Mezey, J. Dnes, L. Bognr, M. Oertel, and Z. Takts, "Real time analysis of brain tissue by direct combination of ultrasonic surgical aspiration and sonic spray mass spectrometry," *Analytical Chemistry*, vol. 83, no. 20, pp. 7729–7735, 2011.

[9] R. J. Goodwin, "Sample preparation for mass spectrometry imaging: Small mistakes can lead to big consequences," *Journal of Proteomics*, vol. 75, no. 16, pp. 4893 – 4911, 2012, special Issue: Imaging Mass Spectrometry: A Users Guide to a New Technique for Biological and Biomedical Research.

[10] A. Limbeck, P. Galler, M. Bonta, G. Bauer, W. Nischkauer, and F. Vanhaecke, "Recent advances in quantitative la-icp-ms analysis: challenges and solutions in the life sciences and environmental chemistry," *Analytical and Bioanalytical Chemistry*, vol. 407, no. 22, pp. 6593–6617, Sep 2015.

[11] M. J. Bailey, N. J. Bright, R. S. Croxton, S. Francese, L. S. Ferguson, S. Hinder, S. Jickells, B. J. Jones, B. N. Jones, S. G. Kazarian, J. J. Ojeda, R. P. Webb, R. Wolstenholme, and S. Bleay, "Chemical characterization of latent fingerprints by matrix-assisted laser desorption ionization, time-of-flight secondary ion mass spectrometry, mega electron volt secondary mass spectrometry, gas chromatography/mass spectrometry, x-ray photoelectron spectroscopy, and attenuated total reflection fourier transform infrared spectroscopic imaging: An intercomparison," *Analytical Chemistry*, vol. 84, no. 20, pp. 8514–8523, 2012.

[12] M. J. Bailey, E. C. Randall, C. Costa, T. L. Salter, A. M. Race, M. de Puit, M. Koeberg, M. Baumert, and J. Bunch, "Analysis of urine, oral fluid and fingerprints by liquid extraction surface analysis coupled to high resolution ms and ms/ms - opportunities for forensic and biomedical science," *Anal. Methods*, vol. 8, pp. 3373–3382, 2016.

[13] S. A. Thomas, D. J. Lloyd, and A. C. Skeldon, "Equation-free analysis of agent-based models and systematic parameter determination," *Physica A: Statistical Mechanics and its Applications*, vol. 464, pp. 27 – 53, 2016.

[14] S. A. Thomas, A. M. Race, R. T. Steven, I. S. Gilmore, and J. Bunch, "Dimensionality reduction of mass spectrometry imaging data using autoencoders," in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, Dec 2016, pp. 1–7.

[15] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[16] L. van der Maaten, "Learning a parametric embedding by preserving local structure," *In Proceedings of the Twelfth International Conference on Artificial Intelligence & Statistics (AI-STATS), JMLR W&CP*, vol. 5, pp. 384–391, 2009.

[17] J. M. Fonville, C. L. Carter, L. Pizarro, R. T. Steven, A. D. Palmer, R. L. Griffiths, P. F. Lalor, J. C. Lindon, J. K. Nicholson, E. Holmes, and J. Bunch, "Hyperspectral visualization of mass spectrometry imaging data," *Analytical Chemistry*, vol. 85, no. 3, pp. 1415–1423, 2013.

[18] W. M. Abdelmoula, K. krkov, B. Balluff, R. J. Carreira, E. A. Tolner, B. P. F. Lelieveldt, L. van der Maaten, H. Morreau, A. M. J. M. van den Maagdenberg, R. M. A. Heeren, L. A. McDonnell, and J. Dijkstra, "Automatic generic registration of mass spectrometry imaging data to histology using nonlinear stochastic embedding," *Analytical Chemistry*, vol. 86, no. 18, pp. 9204–9211, 2014.

[19] W. M. Abdelmoula, R. J. Carreira, R. Shyti, B. Balluff, R. J. M. van Zeijl, E. A. Tolner, B. F. P. Lelieveldt, A. M. J. M. van den Maagdenberg, L. A. McDonnell, and J. Dijkstra, "Automatic registration of mass spectrometry imaging data sets to the allen brain atlas," *Analytical Chemistry*, vol. 86, no. 8, pp. 3947–3954, 2014.

[20] W. M. Abdelmoula, B. Balluff, S. Englert, J. Dijkstra, M. J. T. Reinders, A. Walch, L. A. McDonnell, and B. P. F. Lelieveldt, "Data-driven identification of prognostic tumor subpopulations using spatially mapped t-sne of mass spectrometry imaging data," *Proceedings of the National Academy of Sciences*, vol. 113, no. 43, pp. 12 244–12 249, 2016.

[21] L. van der Maaten, "Barnes-Hut-SNE," *CoRR*, vol. abs/1301.3342, 2013. [Online]. Available: http://arxiv.org/abs/1301.3342

[22] M. Stoeckli, D. Staab, and A. Schweitzer, "Compound and metabolite distribution measured by MALDI mass spectrometric imaging in whole-body tissue sections," *International Journal of Mass Spectrometry*, vol. 260, no. 23, pp. 195 – 202, 2007, imaging Mass Spectrometry Special Issue.

[23] D. S. Cornett, S. L. Frappier, and R. M. Caprioli, "MALDI-FTICR imaging mass spectrometry of drugs and metabolites in tissue," *Analytical Chemistry*, vol. 80, no. 14, pp. 5648–5653, 2008.

[24] C. L. Carter, C. W. McLeod, and J. Bunch, "Imaging of phospholipids in formalin fixed rat brain sections by matrix assisted laser desorption/ionization mass spectrometry," *Journal of The American Society for Mass Spectrometry*, vol. 22, no. 11, pp. 1991–1998, 2011.

[25] A. D. Palmer, R. Griffiths, I. Styles, E. Claridge, A. Calcagni, and J. Bunch, "Sucrose cryo-protection facilitates imaging of whole eye sections by MALDI mass spectrometry," *Journal of Mass Spectrometry*, vol. 47, no. 2, pp. 237–241, 2012.

[26] M. K. Passarelli and N. Winograd, "Lipid imaging with time-of-flight secondary ion mass spectrometry (tof-sims)," *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, vol. 1811, no. 11, pp. 976 – 990, 2011, lipidomics and Imaging Mass Spectrometry.

[27] I. M. Taban, A. F. M. Altelaar, Y. E. M. van der Burgt, L. A. McDonnell, R. M. A. Heeren, J. Fuchser, and G. Baykut, "Imaging of peptides in the rat brain using MALDI-FTICR mass spectrometry," *Journal of the American Society for Mass Spectrometry*, vol. 18, no. 1, pp. 145–151, 2007.

[28] S. Khatib-Shahidi, M. Andersson, J. L. Herman, T. A. Gillespie, and R. M. Caprioli, "Direct molecular analysis of whole-body animal tissue sections by imaging MALDI mass spectrometry," *Analytical Chemistry*, vol. 78, no. 18, pp. 6448–6456, 2006.

[29] F. M. Green, F. Kollmer, E. Niehuis, I. S. Gilmore, and M. P. Seah, "Imaging g-sims: a novel bismuth-manganese source emitter," *Rapid Communications in Mass Spectrometry*, vol. 22, no. 16, pp. 2602–2608, 2008.

[30] M. K. Passarelli and A. G. Ewing, "Single-cell imaging mass spectrometry," *Current Opinion in Chemical Biology*, vol. 17, no. 5, pp. 854 – 859, 2013, in vivo chemistry Analytical techniques.

[31] M. K. Passarelli, C. F. Newman, P. S. Marshall, A. West, I. S. Gilmore, J. Bunch, M. R. Alexander, and C. T. Dollery, "Single-cell analysis: Visualizing pharmaceutical and metabolite uptake in cells with label-free 3d mass spectrometry imaging," *Analytical Chemistry*, vol. 87, no. 13, pp. 6696–6702, 2015.

[32] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: a review of classification and combining techniques," *Artificial Intelligence Review*, vol. 26, no. 3, pp. 159–190, Nov 2006.

[33] F. Lotte, M. Congedo, A. Lcuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for eeg-based braincomputer interfaces," *Journal of Neural Engineering*, vol. 4, no. 2, p. R1, 2007.

[34] T.-S. Lim, W.-Y. Loh, and Y.-S. Shih, "A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms," *Machine Learning*, vol. 40, no. 3, pp. 203–228, Sep 2000.

[35] B. Wu, A. T, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao, "Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data," *Bioinformatics*, vol. 19, pp. 1636–1643, 2003.

[36] M. Wagner, D. Naik, and A. Pothen, "Protocols for disease classification from mass spectrometry data," *PROTEOMICS*, vol. 3, no. 9, pp. 1692–1698, 2003.

[37] T. Nguyen, S. Nahavandi, D. Creighton, and A. Khosravi, "Mass spectrometry cancer data classification using wavelets and genetic algorithm," *FEBS Letters*, vol. 589, no. 24, pp. 3879 – 3886, 2015.

[38] S. Gu, R. Cheng, and Y. Jin, "Feature selection for high-dimensional classification using a competitive swarm optimizer," *Soft Computing*, Oct 2016.