

## CANCER DETECTION IN MASS SPECTROMETRY IMAGING DATA BY RECURRENT NEURAL NETWORKS

*F. Ghazvinian Zanjani*<sup>1</sup>, *A. Panteli*<sup>1</sup>, *S. Zinger*<sup>1</sup>, *F. van der Sommen*<sup>1</sup>, *T. Tan*<sup>1</sup>  
*B. Balluff*<sup>2</sup>, *D. R. N. Vos*<sup>2</sup>, *S. R. Ellis*<sup>2</sup>, *R. M. A. Heeren*<sup>2</sup>, *M. Lucas*<sup>3</sup>, *H. A. Marquering*<sup>3</sup>  
*I. Jansen*<sup>3</sup>, *C. D. Savci-Heijink*<sup>3</sup>, *D. M. de Bruin*<sup>3</sup> and *P. H. N. de With*<sup>1</sup>

<sup>1</sup> Eindhoven University of Technology, SPS-VCA, 5612 AJ Eindhoven, The Netherlands

<sup>2</sup> Maastricht Multimodal Molecular Imaging Institute, University of Maastricht, The Netherlands

<sup>3</sup> Biomedical Eng. & Physics, Amsterdam UMC, University of Amsterdam, The Netherlands

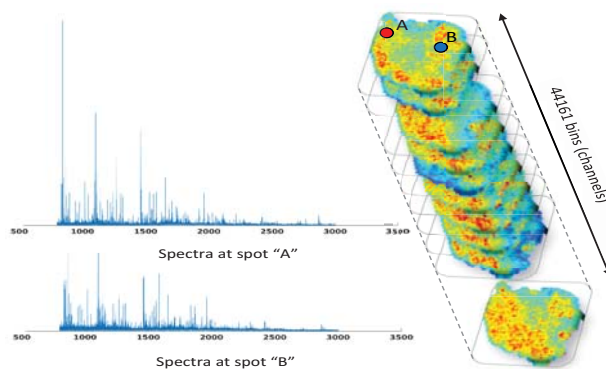
### ABSTRACT

Mass spectrometry imaging (MSI) reveals the localization of a broad scale of compounds ranging from metabolites to proteins in biological tissues. This makes MSI an attractive tool in biomedical research for studying diseases. Computer-aided diagnosis (CAD) systems facilitate the analysis of the molecular profile in tumor tissues to provide a distinctive fingerprint for finding biomarkers. In this paper, the performance of recurrent neural networks (RNNs) is studied on MSI data to exploit their learning capabilities for finding irregular patterns and dependencies in sequential data. In order to design a better CAD model for tumor detection/classification, several configurations of Long Short-Time Memory (LSTM) are examined. The proposed model consists of a 2-layer bi-directional LSTM, each containing 100 LSTM units. The proposed RNN model outperforms the state-of-the-art CNN model by 1.87% and 1.45% higher accuracy in mass spectra classification on lung and bladder cancer datasets with a sixfold faster training time.

**Index Terms**— Recurrent Neural Networks (RNN), Mass Spectrometry Imaging (MSI), cancer detection, deep learning, long short-term memory (LSTM).

### 1. INTRODUCTION

Cancer is a devastating disease that affects millions of people worldwide per year and detecting cancerous regions is very important for diagnosis and treatment [1]. Mass Spectrometry (MS) is an analytical method for measuring the masses of particles and can be applied to analyze tumors [2]. MS data includes one-dimensional ion-particle signals as a function of the mass-to-charge ratio. Mass spectrometry imaging (MSI) is a novel molecular imaging technique that measures the MS data in several spots on the surface of the tissue, usually applied on a 2D grid which can be visualized as a multi-channel image [2]. Such an image represents the spatial distribution of proteins, peptides, lipids and other small



**Fig. 1:** (left) spectra in two locations on the bladder-tissue surface; (right) multi-channel MS image

molecules within the surface of the tissue slice. This complex signal has high potential for addressing different diagnostic and prognostic clinical studies, like detecting tumorous and healthy regions, classifying cancer subtypes, grading metastases, etc. However, analysis and interpretation of MSI data requires specialized methods. Figure 1 shows an MSI example from a slice of bladder tissue, which illustrates the need for computer-aided diagnosis (CAD) of such a signal.

Recent advances in machine learning have made artificial neural networks (ANNs) a powerful tool for processing the complex high-dimensional data as an image. Recurrent Neural Networks (RNNs) are a family of ANNs, initially proposed for learning temporal dynamics in sequential data, where there are unknown dependencies between the elements of a sequence [3]. The recurrent properties of RNNs have proven useful in multi-dimensional sequence processing and irregular pattern extraction and are able to learn spatial dependencies in signals such as MSI data [3]. Long short-time memory (LSTM) networks are a subset of RNNs and address the problem of exploding and vanishing gradient in conventional RNNs [4]. LSTM networks are especially capable in handling the learning of long-term dependencies and apply-

ing LSTM networks can further increase the performance of present gold-standard cancer identification methods [5, 6].

## 2. RELATED WORK

Bright-field microscopy is a common practice in clinical pathology. MSI has several advantages over the conventional histopathology bright-field microscopy [7]. However, interpreting and diagnosing the mass spectra is not so straightforward as with stained histopathological slides (e.g. by pathologists), thus making the CAD approach essential. Several studies investigated MSI signal processing and pattern recognition with machine learning methods for feature extraction and classification [8–11]. Apart from the standard signal processing techniques, more advanced methods including ANNs like Convolutional Neural Networks (CNNs), are investigated. For example, in the work of Spencer *et al.* [12], auto-encoders are used to reduce the dimensionality of the MSI data and identify only the most indicative variables. In [11] and [13], the CNNs for biological tissue classification, such as lung, pancreatic and gastric cancer, are investigated.

In this paper, the performance of LSTM networks, in combination with various configurations, is studied for classifying mass spectra applied to clinical data, containing tumorous and healthy tissues. Two recent state-of-the-art developments are considered for comparison with the proposed model: (1) the PCA/LDA method by Boskamp *et al.* [8] that uses principal components analysis (PCA) and linear discriminant analysis (LDA) methods, and (2) the CNN approach by Behrmann *et al.* [11] that uses the recent advanced techniques in a deep CNN called *IsotopeNet*, to classify cancerous lung tissue.

The contribution of this work are twofold. First, different LSTM architectures are explored to maximize the classification performance on MS data. Several configurations have been examined such as increasing complexity (e.g. by adding hidden layers), implementing a bi-directional architecture, applying Lasso regression (or L1 regularization) and augmenting the MS data. Second, the performance of the LSTM approach is compared with the PCA/LDA and *IsotopeNet* models. With experiments, it is shown that LSTM networks are able to learn the local and non-local patterns in MS data and outperform the two state-of-the-art methods in mass spectra classification.

## 3. METHODOLOGY

### 3.1. LSTM network architecture

A variation of the LSTM architecture proposed by Zhang *et al.* [14] is considered as the baseline LSTM model. This model consists of a single LSTM layer with 500 LSTM units, followed by a dense layer with two output nodes and the rectified linear unit (ReLU) non-linearity function. The input to the

network is all mass spectra of a sample (e.g. a vector with 27,286 elements) and the output is the binary label representing the input class.

For achieving a better performance on MS-data classification, several variants to the baseline model are investigated and their impacts on the classification performance are evaluated:

1. The number of units in the LSTM layer were varied from 10 up until 1,000. Adding more LSTM units increases the dimensionality of the latent space and leads to increased learning capacity of the network.
2. The number of layers (depth) of the model was varied from 1-4 layers deep. Exploiting more layers helps the network learn higher-order dependencies and consequently capturing more complex patterns in a hierarchical structure.
3. With and without bi-directional architecture. For merging the forward and backward networks averaging, concatenation and summation operators are used.
4. Batch normalization and dropout, on the weights of the input layer, were introduced. Applying dropout can decrease over-fitting whilst batch normalization can speed up the training time [15–17].
5. L1 regularization, or Lasso regression, is used for dimensional selection due to the sparsity of the MS sequences, in order to further increase the performance of the model [18].
6. Data augmentation is proposed for increasing the variation in the data and amplifying the generalization power of the model. However, applying it on the MSI data has no established precedent and is not straightforward. A variation of the PCA data augmentation method called *Fancy-PCA* was employed, which is adjusted from the work of Krizhevsky *et al.* [19]. To do so, first the covariance matrix of training data is computed. Afterward, by applying the Singular Value Decomposition, eigenvalues are augmented by a random factor in the range of [1, 1.7] and the new samples are generated.

Combinations of all above considerations are proposed for improving the baseline model and achieving a higher performance for MSI data classification.

### 3.2. Experiments

#### 3.2.1. Datasets

Two datasets were used to evaluate the proposed method. The first dataset is used in the work of Boskamp *et al.* [8] and Behrmann *et al.* [11] which is publicly available. This dataset includes the MS data from two types of lung cancer: carcinoma and squamous cell carcinoma. For better comparison of the acquired results with [8] and [11], the same data partitioning was used. The data was divided into 8 sets and was used with a fourfold cross-validation scheme. The data consists of a total of 4,672 MS samples from 8 patients [8].

The second dataset in this paper is a bladder-tissue dataset from 9 patients. The tissues were serially sectioned by the Academic Medical Center (AMC) in Amsterdam. Half of the provided sections (e.g. all odd-numbered sections) are used for hematoxylin and eosin (H&E) staining and bright-field microscopy. The histological annotations were provided by uropathologist and were used as the ground truth for the other half of the sections (all even-numbered sections). In turn, the last are used for the MSI at Maastricht University (Maastricht, The Netherlands). The MS data were obtained at a spatial resolution of  $50\mu\text{m}$  with a RapifleX Time-of-Flight mass spectrometer across the mass-to-charge-ratio ( $m/z$ ) range of  $800 - 3,000Da$  (see Figure 1). The resolution is  $0.0498Da$ , resulting in 44,161 bins over the whole range. A subset of 13,000 samples are used, all representing either urothelial cell carcinoma or healthy urothelium and healthy detrusor muscle tissue. The samples are distributed evenly over those two classes. The dataset is divided into a threefold cross-validation scheme at patient level. Here, the task is the classification of mass spectra belonging to tumorous and muscular tissue. Figure 2 shows an H&E-stained tissue with the overlaid ground truth.

### 3.2.2. Metrics

The performance of the binary classification is measured with the balanced accuracy metric, calculated by  $\frac{1}{2}(\frac{TP}{P} + \frac{TN}{N})$ ; where  $P$  and  $N$  are the numbers of positive and negative predictions, respectively. Here,  $TP$  and  $TN$  are the numbers of true positives and true negatives, respectively. Balanced accuracy is more indicative than accuracy, as it considers the possible effect of an imbalanced dataset [20].

For exploring the best network architecture, all experiments are performed on the lung cancer dataset only. Thereafter, the best model is selected and its performance is evaluated on the bladder dataset. This is done to ensure that the performance of the suggested model is not data-dependent. Subsequently, a comparison is made between the LSTM approach and the two state-of-the-art methods (i.e. PCA/LDA and IsotopeNet). This experimental setup prevents possible bias towards the training data. As complementary metrics, the F-1 score and the area under the curve (AUC) of the best architecture are reported.

## 4. RESULTS

In Table 1, the performance of the baseline model along with several configurations of the LSTM network are reported. For training all variants of the model, the following constant configuration was applied: batch size of 32, learning rate of  $10e-4$ , maximum number of 100 epochs, RMSprop optimizer, binary cross-entropy loss, and the use of *HE* parameters initialization [21]. The observations of the experimental results are as follows.

- *Increasing the number of LSTM units* in a single-layer

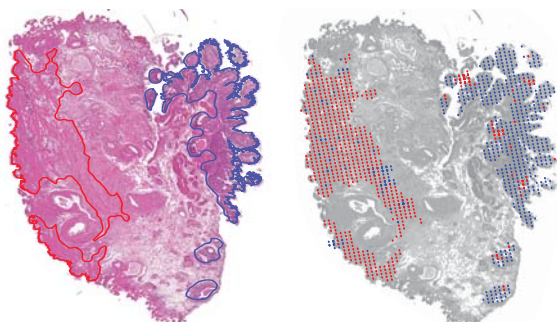
**Table 1:** Exploration results of experiments

Exploration parameters deviating from the base architecture	Balanced accuracy	Training time[s]
(Base) 1 layer, 100 LSTM units	0.8278	1170.8
10 LSTM units	0.8044	964.3
20 LSTM units	0.8215	1290.7
50 LSTM units	0.8296	1534.2
200 LSTM units	0.8304	1213.4
500 LSTM units	0.8308	1683.1
1000 LSTM units	0.8381	2291.6
2 layers	0.8295	460.0
3 layers	0.8281	397.7
4 layers	0.8287	370.8
Bidirectional, concat. merging	0.8301	1610.0
Bidirectional, averaging merging	0.8359	1652.7
Bidirectional, summing merging	0.8289	1722.9
2 Layers, Batch normalization	0.8180	537.0
2 Layers, Dropout (0.5 dropout rate)	0.8350	1367.0
L1 regularization (before Base)	0.8479	220.4
Data augmentation (x2 signal)	0.8282	2432.2
Data augmentation (x3 signal)	0.8265	3706.7
Data augmentation (x4 signal)	0.8278	6311.6
Batch normalization and dropout	0.8302	904.2
L1 reg., Batch norm. and dropout	0.8415	979.1
(2x) L1 reg., Batch norm., dropout	0.8501	1095.4

network leads to increasing the classification performance. However, using more than 50 units shows no noticeable improvement, only the training/execution time is increased. For example, employing 1,000 LSTM units shows the best performance: the balanced accuracy is improved only about  $7 \times 10^{-3}$  in comparison with 500 units, at the cost of a 48.9% increase in training time.

- *Increasing the depth of network* by adding more hidden LSTM layers, does not increase the accuracy further after 2 layers. However, the training time decreases dramatically (around 60% from the baseline).
- *The bi-directional architecture* by an ensemble of two networks (forward and backward), improves the performance as expected by 1% (for averaging merging) at the cost of a 29.1% increase in training time.
- *Batch normalization* decreases the accuracy by 1.2%, but it reduces the training time by 54.1%.
- *The dropout* technique increases the accuracy by 0.9%, at the cost of a 14.4% higher training time.
- *Data augmentation* does not increase the performance.

Table 2 shows the best performing architecture. It consists of two bi-directional LSTM networks, with 100 LSTM units, in its hidden layers and the L1 regularization. The dense layer maps the input mass spectra into 100-dimensional space on which the LSTM units are applied. The L1 regularization is applied to the weights of the dense layer, in order to force some weights towards zero. The bi-directional network uses average merging to combine the output of the LSTM layers



**Fig. 2:** Prediction of the proposed model on bladder tissue; (left) H&E slide with the overlaid ground truth; (right) model predictions at the laser spots of sampled MSI data in MSI.

**Table 2:** Best performing RNN LSTM architecture

Layer	Shape
Input*	(-, 1, 27286)
Dense (L1 regularized)	(-, 1, 100)
Bidirectional ("averaging" merging)	(-, 1, 100)
Bidirectional ("averaging" merging)	(-, 100)
Dense (with Softmax activation)	(-, 2)

\* for the bladder data the input shape is (-, 1, 44,161)

**Table 3:** Comparison of CNN and RNN architectures

Method	Training time per epoch [s]	No. of param.
IsotopeNet	12.751	<b>13,935</b>
RNN	<b>2.001</b>	3,050,502

in forward and backward directions. Finally, a dense output layer, followed by a softmax activation layer, maps this embedded space into the output labels. Dropout did not contribute to the final architecture as it does not improve the performance on the lung dataset. However, when training the model on the bladder data, dropout was added. This was necessary to avoid over-fitting on the bladder mass spectra because of its higher dimensionality equal to 44,161 (about 38.2% more than the lung dataset). Figure 2 illustrates the performance of the RNN architecture on the bladder dataset.

The balanced accuracy, AUC and F-1 score results of the proposed LSTM network, in comparison with the PCA/LDA and IsotopeNet are shown in Table 4. The PCA/LDA was implemented according to work of Boskamp *et al.* [8] for up to 100 principal components. The RNN and IsotopeNet approaches use an interquartile range for the dispersion of the results. Furthermore, Table 3 shows the measured training time of the RNN and the CNN architecture in TensorFlow on a GTX-1080 graphics card.

## 5. DISCUSSION

The LSTM approach obtains on the lung cancer dataset a 7.68% higher accuracy than the conventional PCA/LDA approach [8]. It also outperforms IsotopeNet, proposed by the recent advances in CNNs, by about 1.87% higher accuracy on the same set. On the bladder dataset, RNN performs 15.7%

**Table 4:** Balanced Accuracy, AUC and F-1 score results

Method	Lung cancer data	Bladder dataset
<i>Balanced Accuracy</i>		
PCA/LDA	0.7869	0.6689
IsotopeNet	0.8450 ± 0.007	0.8115 ± 0.1634 [22]
RNN	<b>0.8637</b> ± 0.012	<b>0.8260</b> ± 0.023
<i>(F-1 score, AUC)</i>		
PCA/LDA	(0.8259, 0.7868)	(0.673, 0.666)
IsotopeNet	(0.853, 0.841)	(0.834, <b>0.893</b> ) [22]
RNN	<b>(0.865, 0.870)</b>	<b>(0.840, 0.832)</b>

and 1.45% higher than PCA/LDA and IsotopeNet, respectively. Although the proposed RNN model has over 200 times more parameters than the IsotopeNet, it trains 6 times faster. This is explained by the lack of convolution operators, as used by CNN, and by the simple forward and backward gradient propagation of an LSTM network.

Adding more than 100 units to a single-layer LSTM network does not further improve the results. This can originate from saturated learning of the network, meaning that there are increasingly less dependencies to be learned. Similarly, increasing the depth without performance improvements can indicate that a single-layer network already has sufficient complexity for learning the task. The reason that data augmentation did not improve the performance may come from the aspect that the added variations to the data did not influence the critical input space close to the decision boundary between two classes.

## 6. CONCLUSIONS

The proposed LSTM model, for cancer detection and classification on MSI data, consists of 2 hidden layers in a bi-directional architecture, which outperforms the recent advanced CNN approach in MS classification by a moderate 1.87% and 1.45% higher accuracy on two clinical datasets, but with a 6 times faster training time. These aspects are highly relevant for an efficient CAD system. The proposed model is motivated by the nature of MSI data containing local and non-local dependencies between the elements of the spectrum, capturing the cancerous fingerprint in data. This finding is in agreement with the presence of biological signatures for isotopes and proteins in mass spectra. The LSTM networks have proven to model well the long irregular dependencies in sequential data for learning the patterns captured by MS data. It is important to mention that the proposed model has similar performance results on both lung and bladder data (the latter was left out during the exploration phase), it shows that the proposed model has high generalization power and is not biased to a specific dataset.

As a future work, to improve the results of RNN, it is advised to use LSTM networks to analyze neighboring spectra located in a 2D region to improve the classification results. In addition, the combination of CNN and RNN can be an alternative approach into further increasing the performance, since their combined properties can complement each other.

## 7. REFERENCES

- [1] American Cancer Society, “Cancer facts & figures 2018,” pp. 3–8, 2018.
- [2] Tyler A Zimmerman, Eric B Monroe, Kevin R Tucker, Stanislav S Rubakhin, and Jonathan V Sweedler, “Imaging of cells and tissues with mass spectrometry: adding chemical information to imaging,” *Methods in cell biology*, vol. 89, pp. 361–390, 2008.
- [3] Roelof K Brouwer, “A method for training recurrent neural networks for classification by building basins of attraction,” *Neural Net.*, vol. 8, no. 4, pp. 597–603, 1995.
- [4] Felix A Gers and E Schmidhuber, “Lstm recurrent networks learn simple context-free and context-sensitive languages,” *IEEE Trans. on Neural Networks*, vol. 12, no. 6, pp. 1333–1340, 2001.
- [5] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals, “Recurrent neural network regularization,” *arXiv preprint arXiv:1409.2329*, 2014.
- [6] Ngoc Hieu Tran, Xianglilan Zhang, Lei Xin, Baozhen Shan, and Ming Li, “De novo peptide sequencing by deep learning,” *Proc. of the National Acad. of Sci.*, vol. 114, no. 31, pp. 8247–8252, 2017.
- [7] Metin N Gurcan and E Boucheron, “Histopathological image analysis: A review,” vol. 2, pp. 147 – 171, 2009.
- [8] Tobias Boskamp and Lachmund, “A new classification method for MALDI imaging mass spectrometry data acquired on formalin-fixed paraffin-embedded tissue samples,” *Biochimica et Biophysica Acta - Proteins and Proteomics*, vol. 1865, no. 7, pp. 916–926, 2017.
- [9] Mark Kriegsmann, Rita Casadonte, and Jörg Kriegsmann et al., “Reliable entity subtyping in non-small cell lung cancer by maldi imaging mass spectrometry on formalin-fixed paraffin-embedded tissue specimens,” *Molecular & Cellular Proteomics*, vol. 15, no. 10, pp. 3081–3089, 2016.
- [10] Walid M. Abdelmoula, Benjamin Balluff, and Sonja Englert et al., “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. 12244–12249, 2016.
- [11] Jens Behrmann, Christian Etmann, Tobias Boskamp, Rita Casadonte, Jörg Kriegsmann, and Peter Maaß, “Deep learning for tumor classification in imaging mass spectrometry,” *Bioinformatics*, vol. 34, no. 7, pp. 1215–1223, 2017.
- [12] Spencer A Thomas, Alan M Race, Rory T Steven, Ian S Gilmore, and Josephine Bunch, “Dimensionality reduction of mass spectrometry imaging data using autoencoders,” in *Computational Intelligence (SSCI), 2016 IEEE Symposium Series on*. IEEE, 2016, pp. 1–7.
- [13] Paolo Inglese, James S McKenzie, and Anna Mroz et al., “Deep learning and 3d-desi imaging reveal the hidden metabolic heterogeneity of cancer,” *Chem. Sci.*, vol. 8, no. 5, pp. 3500–3511, 2017.
- [14] Jinlei Zhang, Junxiu Liu, Yuling Luo, Qiang Fu, Jinjie Bi, Senhui Qiu, Yi Cao, and Xuemei Ding, “Chemical substance classification using long short-term memory recurrent neural network,” in *17th IEEE Int. Conf. on Communication Tech. (ICCT)*, 2017, pp. 1994–1997.
- [15] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi, “Xnor-net: Imagenet classification using binary convolutional neural networks,” in *European Conf. on Computer Vision*. Springer, 2016, pp. 525–542.
- [16] Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio, “How to construct deep recurrent neural networks,” *arXiv preprint arXiv:1312.6026*, 2013.
- [17] Sergey Ioffe and Christian Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proc. of the 32nd Int. Conf. on Mach. Learning*. 2015, vol. 37 of *Proc. of Mach. Learning Research*, pp. 448–456, PMLR.
- [18] Robert Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Stat. Society. Series B (Methodological)*, pp. 267–288, 1996.
- [19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Inf. Proc. Sys.*, 2012, pp. 1097–1105.
- [20] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann, “The balanced accuracy and its posterior distribution,” in *20th Int. Conf. on Pattern Recognition*. IEEE, 2010, pp. 3121–3124.
- [21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in *Proceedings of the IEEE Int. Conf. on computer vision*, 2015, pp. 1026–1034.
- [22] Jannis van Kersbergen, Farhad G. Zanjani, and Svitlana Zinger et al., “Cancer detection in mass spectrometry imaging data by dilated convolutional neural networks,” *Medical Imaging 2018 (Proc. of SPIE)*, No. 10956, (In press).