Analyzing Neural Network Architecture for

Classification of High-Performance Liquid

Chromatography Chromatograms

**Akhil Mandalapu, Max Franke, Bhagyalakshmi Patchipalu, Philemon L. Ramontal, Sean Mondesire**

St. Thomas University

Miami, FL, USA  
mandalapu.akhil@gmail.com, mfranke@stu.edu, bpatchipalu@stu.edu, pramontal@stu.edu,smondesire@stu.edu

Abstract

The combination of high-performance mass spectrometry liquid chromatography (MS-HPLC) with metabolomics is a chemical analytics technique used to measure cancerous and normal cellular activity in a subject. Though the technique is highly accurate and standard, the cancerous-benign classification is traditionally performed by hand with a trained technician visually inspecting each image. The process is time consuming, tedious, and prone to human error. Therefore, to improve the classification procedure, a well-studied, understood, and accurate automated process is needed.

The proposed study explores automating MS-HPLC classification with the use of deep learning artificial neural networks. Particularly, this work determines the optimal network structure for the image classifier, performs statistical analysis on hidden layer depth and the number of neurons per layer, and evaluates image denoising and filtering techniques to yield higher accuracy. This study’s results have yielded a classification accuracy of 92.94% and could lead to further research to fine-tune the MS-HPLC with metabolomics automation process.

Introduction*[[1]](#footnote-1)*

The field of *metabolomics* is a study of chemical processes involving *metabolites*, intermediary compounds in metabolic processes within cells. Metabolomics allows for insight regarding patient health at the cellular level, including the detection of cancerous cells (Schmidt 2004). Because metabolomics requires analysis of high volumes and a variety of metabolites simultaneously, methods such as *Mass Spectrometry* (*MS*), are used for this type of classification (Schmidt 2004).

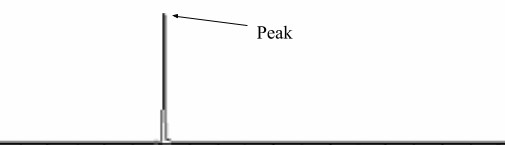
A key technique used in metabolomics couples *Mass Spectrometry* and *Chromatography*, an analytical chemistry technique in which compounds of a sample are separated, detected, and quantified for classification. A type of liquid chromatography, *high-performance liquid chromatography-mass spectrometry* (*HPLC-MS*), is a commonly effective method within metabolomics research (Schmidt 2004).

Unfortunately, techniques involved in the quantification of compounds of interest in metabolomics within cancer research and HPLC-MS are time-consuming and require trained human supervision. Though recent advances in artificial intelligence and computer processing capability have resulted in the widespread adoption of machine learning algorithms, machine learning algorithms have been underexplored as an automated solution to cancer detection.

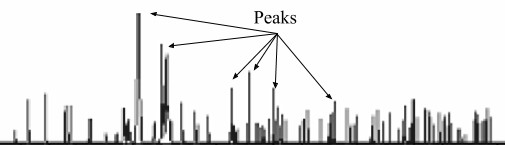
In recent years, *artificial neural networks* (*ANNs*) have become especially popular as *deep learning networks* (*DLNs*) have been shown to outperform other machine learning algorithms in a variety of complex image classification tasks (Raway and Wang 2017), (Guresent and Kayakutlu 2011). However, machine learning classifiers, particularly, ANNs, have yet to be fully explored for HPLC-MS classification. Therefore, this work examines the effectiveness of using ANNs for peak identification. Specifically, this work evaluates the classification accuracy of several convolutional ANN structures for labeling chromatograms as being suitable (containing one peak) or unsuitable (multiple peaks). With this study, the metabolomics community can begin to determine the feasibly of deploying an automated tool for chromatogram image classification. This preliminary study is motivated by the potential of automated HPLC-MS peak detection and classification, which could result in less resource-intensive cancer detection. Improvements in the workflow of metabolomics researchers could lead to more proteins and genes of interest for future cancer research and a more comprehensive understanding of cancer’s influence on the body.

Background

One effective technique used within HPLC-MS analysis is the integration of the spikes or “peaks” of a chromatogram produced from a Mass Spectrometer machine. Chromatograms with a single distinct peak, as shown in Fig. 1(a) are suitable for integration, while those with plenty of peaks of various sizes throughout the chromatogram are not suitable, as shown in Fig 1(b). Traditionally, this filtering process is done manually, is time-consuming, and is resource-intensive. Various methods of automation have been explored utilizing statistical, mathematical, and machine learning models, though for HPLC-MS analysis these methods have not been overwhelmingly successful.



(a)



(b)

Fig. 1: (a) An example of a suitable chromatogram for use in cancer metabolomics analysis; (b) an example of an unsuitable chromatogram.

Chromatography, coupled with Mass Spectrometry, is a more precise method compared to Direct MS, another standard form of detection (Lei, Huhnan, and Sumner 2011). HPLC-MS is suitable for metabolomics due to offering more versatility in the compounds quantified compared to other chromatography coupled methods. Corroborated studies determined that HPLC-MS offers the advantages of significantly reducing analysis time in comparison to traditional column chromatography methods as well as having overall higher resolution and heightened precision and sensitivity (Lakshmi 2015), (Bird 1989). The issue with HPLC-MS resides in its limitations: the cost of equipment, technician training, and dependence on MS machinery.

HPLC-MS is not the only method for peak detection. One attempted automated approach implemented a mathematical algorithm that was validated on a set of Fourier Transform Ion Cyclotron Resonance MS data (Chiron et al.). The algorithmic approach achieved similar results as other techniques, such as the Cadzow approach, but with drastically lower computational costs (Chiron et al. 2014). Computational costs and time for image processing are the biggest limitations to these algorithmic approaches. To counter these issues and reduce computational complexity, denoising techniques have been used instead of peak classification, and a lack of focus on HPLC-MS data, a recurring limitation throughout the literature (Chiron et al. 2014). Another study details a statistical-based algorithm for peak detection of two-dimensional gas chromatography MS data on par with commercially available options. As noted in the previous study, limited computational resources when developing their algorithm led to limitations in the accuracy of the final algorithm (Kim et al. 2014).

Machine learning methods for use in the liquid-chromatography analysis were also developed. In one instance, end-to-end workflow for LC-MS data analysis using machine learning methods and peak merging was developed (Albert 2014). Peak merging with *Support Vector Machines* (*SVM*) and *Partial Least Square-Discriminant Analysis* (*PLS-DA*) resulted in approximately a 14-18% improvement in classifier performance (Albert 2014).

Literature within the field of machine learning has focused on the comparison of classifiers within a variety of classification tasks, namely, image classification and waveform data classification. One study compared classifiers on Electrocardiogram data classification (Patro and Kumar 2017). *Electrocardiogram* (*ECG*) data as a type of waveform data present similar challenges for classification as HPLC-MS data, and thus, the study reveals insights relevant to HPLC-MS analysis. In the study, data was preprocessed by denoising, extracted features of interest based on P-QRS-T fragments from the ECG data, and trained a set of classifiers, ANN, SVMs, and *K-Nearest Neighbor* (*KNN*) (Patro and Kumar 2017). Results of the study include a 93.7% accuracy achieved by the SVM and 92.7% and 92.4% achieved by KNN and ANN respectively (Patro and Kumar 2017). Key findings of the paper include the support of SVMs in the classification of ECG data but with ANN and KNN achieving similar accuracy results and the emphasis on denoising assisting with classification accuracy.

Another perspective that has been explored in the field of machine learning is the capabilities of deep learning networks for image classification. *Deep Convolutional Neural Networks* (*DCNN*) have had various levels of success in solving complex image classification tasks (Rawat and Wang 2017). Figure 2 depicts a CNN, where the input layers accept image data and the hidden layers map those inputs to a classification result in the output layer. Currently, DCNNs are the standard on image classification tasks and, on single label tasks, exceeded human performance. However, the challenges that have arisen with DCNNs such as a gap of understanding of the internal structure, inability to be downscaled to lower-end devices due to their computational resource cost and current gap in unsupervised learning in that most neural networks followed a supervised learning model were also highlighted (Rawat and Wang 2017). Overall, the literature exploring machine learning algorithms has compared classifiers on waveform-like data like HPLC-MS but not specifically on HPLC-MS data.

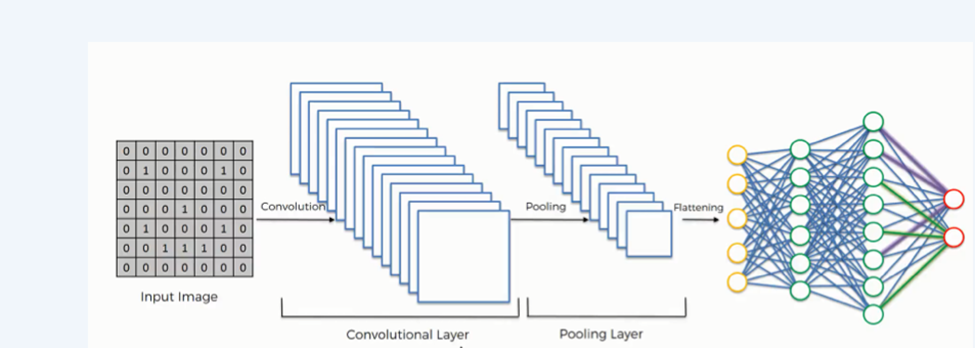


Figure 2:  A Convolutional Neural Network architecture

Methodology

This study examines the accuracy of convolutional neural networks in HPLC-MS peak classification with 64 hidden layers. The networks classify images as being a “*suitable*” chromatogram (one peak) or “*unsuitable*” (multiple peaks).

The networks are trained on a random split of images, 80% delegated for the training set and 20% for the testing set from the total of 10,000 pre-labeled images. 50 instances of each network are being trained and their accuracies for classifying the testing set will be recorded. 64 hidden layers are chosen as it allows for comparison to be made between groups. This experimental design was chosen to determine the relationship between accuracy and increased layers in the ANN.

1. **Participants**

The images used for the training and testing set are HPLC-MS chromatograms created by an ABSCIEX MS machine and viewed through AB Sciex Multiquant software from a metabolomics lab. The ABSCIEX MS machine is designed for use in cancer metabolomics analysis and allows for a large quantity of data to be generated for use in training machine learning algorithms. The AB Sciex Machine provided the appropriate kind of chromatography analysis pertinent to cancer metabolomics as well as access to a large breadth of data to train machine learning algorithms.

The machine learning networks were developed using the Keras Platform and written in the R programming language. Keras offers data regarding the accuracy and performance of the algorithm and for quick development and deployment, allowing it to be an accessible platform for metabolomics researchers to implement easily. R is one of the standard platforms for machine learning as well as statistical analysis and provides concise code and inbuilt matrix operations and statistical tools.

1. **Materials**

Materials needed for this experiment include an AB Sciex 5600 Triple TOF MS machine, AB Sciex Multiquant Software, a desktop computer with a 3.2 GHz Intel Core i5 processor and 16 GB 1600 MHz DDR3 memory configuration and R (version 3.5.3), RStudio (version 1.1.463) and Keras (version 2.2.4) installed on the computer. The instrument is used to collect chromatogram data, and chromatograms are viewed in the AB Sciex Multiquant software. A total of 10,000 chromatogram images with a resolution of 230 x 70 pixels were collected from the AB Sciex software and labeled either “usable” or “not usable” for use in metabolomics analysis. Figures 1 (a) and (b) are samples from the image set. Only a total of 10,000 images were collected due to limitations in access to data.

1. **Procedure**

A set of image data of chromatograms was gathered from an ABSCIEX 5600 Triple TOF MS machine through the AB Sciex software. Through normalization process all images were scaled to 1/255. In order to perform a thorough analysis all input images should be scaled to same size. The Input images were split randomly into training and testing sets with ratio of 80% train and 20% test. The images were also randomly flipped horizontally in order to mimic chromatograms with early peaks as well as to prevent the algorithm from misclassifying based on the location of the peak within the chromatogram rather than the peak itself. The batch size was set to 900, meaning that the algorithm ran through the 10,000 images 900 at a time. A batch size of 900 was selected as it was the maximum number of images that the computer running the algorithm could handle at a time without significant drops in computer performance. The number of epochs each algorithm ran for was 50. There was not enough data for early stopping methods to be used such as stopping training at maximum accuracy. 50 epochs were chosen as it is sufficiently high enough that the accuracy achieved relatively small changes in between epochs.

The kernel size for each layer was set to 5x5 as the criteria that distinguished suitable and not suitable chromatograms were not localized and a 5x5-kernel size was chosen as a balance of reflecting the non-localized feature. The activation function on the first hidden layer on all algorithms was the *Rectified Linear Units* (*ReLU*) activation function. The ReLU activation function was used over the Sigmoid, and tanh activation functions as ReLU allows for the network to learn faster and avoids issues with vanishing gradients in that, more and more steps are required to achieve relative minimum when utilizing gradient descent to train the neural network (Krizhevsky, Sutskever, and Hin-ton 2017)(Sharma 2019). Vanishing gradients require more time in order to train and computational resources while not providing much improvement in the performance of the algorithm, which could waste limited resources, especially in a metabolomics lab context. For 64 hidden layer algorithms, additional hidden layers used a Leaky ReLU activation function, which accounts for the issues faced with a normal ReLU function in that neurons in the network often die. The use of ReLU in the initial hidden layer provides the benefits of normal ReLU when neurons deactivating is not a significant issue compared to later hidden layers. Each layer also had a dropout rate of ’0.25’ in order to prevent overfitting by removing some of the inputs from some of the neurons (Srivastava et al. 2014). The dropout rate was arbitrarily chosen as an intermediate value between 0.5 and 0.1 (Kwak and Park 2016).

A single pooling layer was in between each hidden convolutional layer with a filter of 2x2. A 2x2 filter is commonly used in order to prevent overfitting, and larger sizes often are too destructive as it reduces the parameters (Karpathy 2019). The final layer was a dense layer that was used in order to reference the entire network, while convolutional layers only refer to local layers. A dense layer was used with a linear activation function in order to reduce the dimensionality and number of parameters in order to be passed onto the output. Finally, an output layer was used with a Softmax activation function in order to output probabilities for potential predictions the algorithm makes, which is particularly useful in the metabolomics classification context as researchers can filter out algorithm predictions based on algorithm confidence (Hamza 2018).

The algorithm was optimized by creating a function that adds first the number of layers and defines the number of neurons per layer to the keras model classifier. A loop was created inside the function, which added layers linear exponentially to provide accuracies for each layer with permutation combinations of 10 trails each. Trials were conducted for each combination of Layers and permutations, which were set to exponential values of 1,2,4,8,16,32,64. Furthermore, Exploratory Analysis and Hypothesis testing were performed to determine if there was any statistically significant differences.

Figure 3 explains a 2-layer neural network architecture with 64 neurons per hidden layer. Images are inputted into the hidden layer and are transformed using an activation function to the next hidden layers.

Results and Discussion

The 1-Layer algorithm achieved a mean accuracy of 90.06% and a standard deviation of 0.0408, while the 2-Layer algorithm achieved a mean accuracy of 92.94 % and standard deviation of 0.0296. 4-Layer algorithm achieved a mean accuracy of 79.94% with a standard deviation of 0.1185. 8-Layer algorithm achieved an accuracy of 81.97% with a standard deviation of 0.1191. 16-Layer algorithm achieved a mean accuracy of 51.16% with a standard deviation of 0.0526. 32-Layer algorithm achieved a mean accuracy of 49.80% with standard deviation of 0.0431, and 64-Layer achieved a mean accuracy of 51.19% with a standard deviation of 0.0403 as shown in Table 1. Additionally, a One-Way Analysis of Variance (ANOVA) test was conducted with a significant P-Value of 2 x 10-16 less than alpha 0.05 and with an F value (502.7) therefore rejecting the null hypothesis and accepting the alternate hypothesis that there is a statistically significant improvement in classifier performance with differing number of hidden layers (Kim 2014).

The difference between average accuracy and the 7 model algorithms varied significantly. The difference between 1 Layer and 2 Layer was only 2% however from 4 Layer down to 64 Layer there were large variations as shown in Table 1. Tukey test was conducted of the 7 layers and determined there was statistical significant differences in layer 4,8,16.

These results refute a hypothesis that an increase in hidden layer count would significantly improve image classification task accuracy. However, there is a slight difference in accuracy between 1 and 2 hidden layers.

TABLE I: Neural Network Layer ANOVA Results (Testing Set)

|  |  |  |  |
| --- | --- | --- | --- |
| Topology | Neurons | Average Accuracy | Standard Deviation |
| 1 Layer | 1,2,4,8,16,32,64 | 0.9006365 | 0.0408582 |
| 2 Layer | 1,2,4,8,16,32,64 | 0.9294477 | 0.0296785 |
| 4 Layer | 1,2,4,8,16,32,64 | 0.7994630 | 0.1185364 |
| 8 Layer | 1,2,4,8,16,32,64 | 0.8197357 | 0.1191383 |
| 16 Layer | 1,2,4,8,16,32,64 | 0.5116038 | 0.0526226 |
| 32 Layer | 1,2,4,8,16,32,64 | 0.4980021 | 0.0431719 |
| 64 Layer | 1,2,4,8,16,32,64 | 0.5119265 | 0.0403582 |

Additionally, it is important to note the smaller deviation of the 2-layer algorithm in comparison to the other layers since models created with 64 layers would have less variation in their accuracy. While the 2-layer algorithm overall achieved a higher accuracy, the single hidden layer algorithm achieved around 92.94% accuracy.

Furthermore, the dataset used to train the algorithms was relatively small at 10000 images and configurations such as dropout, which are used in order to prevent overfitting, were kept constant between the algorithms. Figure 4 demonstrates a boxplot of the mean accuracies between each layer and permutation combination. In Figure 5, the average accuracy takes a drastic drop off in 16 layer. It is possible that overfitting occurred in 16 layer. Further research would need to be conducted. Overall, the experimental data refute the hypothesis that an increase in hidden layers would achieve higher accuracy than single and double hidden layer algorithms. Instead, the data suggest that the less complex 2-hidden layer algorithm results in the highest overall accuracy.

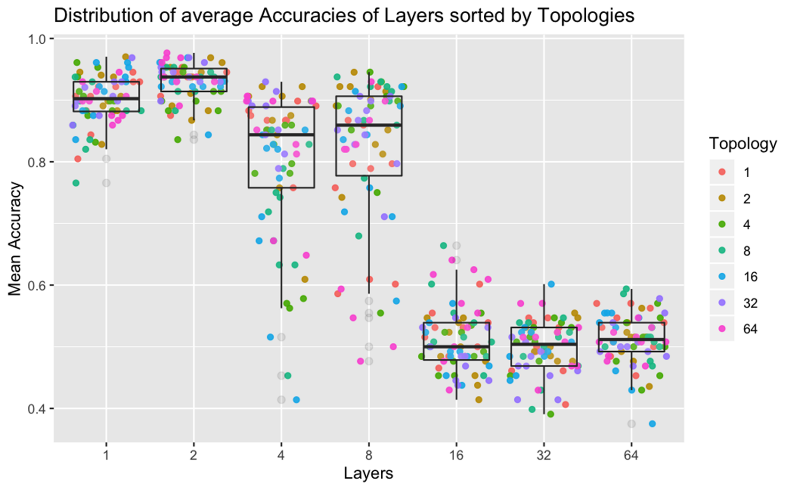


Figure 4: Distribution of each layer with the neurons as points per layer. The variance for a 4-layer network is visually bigger than for a 2-layer network.

Figure 5: Average Accuracies of each layer and neuron combinations.

Conclusion

The study evaluated the effectiveness and feasibility of using convolutional neural networks to classify chromatograms as suitable or unsuitable. The experiments built models, with 64 hidden layers. The neural network with 2 hidden convolutional layers achieved the highest average accuracy of 92.94% in comparison to 90.06% with a 1-hidden layer network. This conclusion supports 2-hidden layer networks as superior in image classification of MS-HPLC data compared to 1, 4, 8, 16, 32, 64 hidden layer neural networks.

Further analyses within the layer combination showed that the number of permutations in the layer networks only showed significant differences in individual cases. The layer networks 1, 2, 8, 32, and 64 do not show any significant difference between the permutation combinations. Only layer-networks 4 and 16 show significant differences in the permutation combinations. The high accuracies associated with these layer counts could allow a neural network to be more accessible for use in laboratories with few computational resources

In conclusion, the high accuracies witnessed in three of the neural network models demonstrated the positive feasibility of automating the HPLC-MS classification. This study could affect the field of metabolomics in attempting to detail the ability of deep learning networks on classifying chromatogram data. Further extensions and examinations of this work’s effectiveness in the field of chemistry, it could lead to the automation of HPLC-MS data and faster analysis of metabolomic data for use in cancer research. This paper is a stepping-stone for future research, which must utilize a larger number of input data as well as compare various configurations such as type of hidden layer, dropout percentages, loss functions, and optimizers.

Acknowledgement

Chromatogram data received from Baylor College of Medicine.

References

Schmidt, C.W. “Metabolomics: what's happening downstream of DNA.” In: *Environmental Health Perspectives* 112.7 (2004). DOI: 10.1289/ehp.112-a410.

Rawat, W. and Wang, Z. “Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review”. In: *Neural Computation* 29.9 (2017). DOI: 10.1162/neco a 00990.

Guresen, E. and Kayakutlu, G. “Definition of artificial neural networks with comparison to other networks”. In: *Procedia Computer Science* 3 (2011), pp. 426–433. DOI: 10.1016/j.procs.2010.12.071.

Dertat, A. *Applied Deep Learning - Part 1: Artificial Neural Networks*. Aug. 2017. URL: https:/ / towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6?gi = f17e2a8389b0.

Patro, K. K. and Kumar, P.R.. “A Machine Learning Classification Approaches for Biometric Recognition System using ECG Signals”. In: *Journal of Engineering Science and Technology Review* 10.6 (2017), pp. 1–8. DOI: 10.25103/jestr.106.01.

Lei, Z., Huhman, D.V., and Sumner, LW. “Mass Spectrometry Strategies in Metabolomics”. In: *Journal of Biological Chemistry* 286.29 (2011), pp. 25435–25442. DOI: 10.1074/jbc.r111.238691.

Lakshmi, S. “A Review on Chromatography with High Performance Liquid Chromatography (HPLC) and its Functions”. In: *Journal of Pharmaceutical Analysis* 4.1 (Mar. 2015). p. 10, pp. 1–15.

Bird., I.M. “High performance liquid chromatography: principles and clinical applications.” In: *Bmj* 299.6702 (1989), pp. 783–787. DOI: 10.1136/bmj.299.6702.783.

Chiron, L. et al. “Efficient denoising algorithms for large experimental datasets and their applications in Fourier transform ion cyclotron resonance mass spectrometry”. In: *Proceedings of the National Academy of Sciences* 111.4 (2014), pp. 1385–1390. DOI: 10 . 1073 / pnas . 1306700111.

Kim, S. et al. “A NEW METHOD OF PEAK DETECTION FOR ANALYSIS OF COMPREHENSIVE TWO-DIMENSIONAL GAS CHROMATOGRAPHY MASS SPECTROMETRY DATA”. In: *The Annals of Applied Statistics* 8.2 (2014), pp. 1209–1231. ISSN: 19326157. URL: http://www.jstor.org/stable/24522093.

Albert, F.F. “Machine Learning Methods for the Analysis of Liquid Chromatography-Mass Spectrometry Datasets in Metabolomics”. PhD thesis. 2014.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. “ImageNet classification with deep convolutional neural networks”. In: *Communications of the ACM* 60.6 (2017), pp. 84–90. DOI: 10.1145/3065386.

Sharma, S. *Activation Functions in Neural Networks*. URL: https://towardsdatascience.com/activationfunctions-neural-networks-1cbd9f8d91d6.

Srivastava, N. et al. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: *Journal of Machine Learning Research 15* (2014), pp. 1929– 1958. URL: http://www.jmlr.org/papers/volume15/ srivastava14a/srivastava14a.pdf.

Kwak, N. and Park, S. *Analysis on the Dropout Effect in Convolutional Neural Networks.* Nov. 2016. URL: https://link.springer.com/chapter/10.1007/978-3-319-54184-6\_12.

Karpathy, A. *Convolutional Neural Networks (CNNs/ConvNets*. URL: https://cs231n.github.io/convolutional-networks/.

Hamza, M. *Softmax Function, Simplified*. Nov. 2018. URL: https://towardsdatascience.com/softmaxfunction-simplified-714068bf8156.

Kim, H.Y.. “Analysis of variance (ANOVA) comparing means of more than two groups”. In: *Restorative Dentistry Endodontics* 39.1 (2014), p. 74. DOI: 10. 5395/rde.2014.39.1.74.

1. Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved. [↑](#footnote-ref-1)