Deep Learning to Classify Single-Cell RNA Sequencing in Primary Glioblastoma

ABSTRACT

Recent advances in single-cell RNA sequencing technologies enable deep insights into cellular development, gene regulation, and phenotypic diversity by measuring gene expression for thousands of cells in a single experiment. This results in high-throughput datasets and requires the development of new types of computational approaches to extract the useful and valuable underlying biological information of individual cells in heterogeneous biological populations. To addresses these approaches, in this paper, we introduce a deep learning technique to classify single cell types data from five primary Glioblastomas. We show that the deep learning method has the ability to correctly infer and classify cell type not used during the training process of the algorithm. Further, the deep learning method has the ability to identify the predictor variable Aquaporin 4 (AQP4), as the most important to make these predictions. Such computational approaches, as those presented in this study will enable researchers to better characterize the intratumoral heterogeneity in primary Glioblastoma.

KEYWORDS

Single cell, deep learning, Glioblastoma

ACM Reference Format:

. 2019. Deep Learning to Classify Single-Cell RNA Sequencing in Primary Glioblastoma. In *EATIS 2020*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Single-cell RNA-sequencing (scRNA-seq) has become increasingly popular for profiling transcriptomic states of individual cells in heterogeneous biological populations [2, 11, 12, 24, 28]. scRNA-seq which profiles the transcriptome of individual cells (as opposed to ensemble of cells) has already led to several new and interesting findings. These include the level of heterogeneity within a population of cells [4], the identification of new markers for specific types of cells [10], and the temporal stages involved in the progression of various developmental processes [23]. It has been used to profile diverse systems including cancer tumors [5, 19], and cell types within the mouse brain [27], amongst others.

Tumor heterogeneity poses a major challenge to cancer diagnosis and treatment. It can manifest as variability between tumors, wherein different stages, genetic lesions or expression programs are associated with distinct outcomes or therapeutic responses [7, 18, 25]. Alternatively, cells from the same tumor may harbor different mutations or exhibit distinct phenotypic or epigenetic states [6, 8, 17, 20]. Such intratumoral heterogeneity is increasingly appreciated as a determinant of treatment failure and disease recurrence [3].

The brain is one of the most complex organs in the human body that works with billions of cells. A brain tumor arises when there is uncontrolled division of cells forming an abnormal group of cells around or inside the brain. Brain tumors can be classified to benign or malignant. Malignant tumors are cancerous and can originate from the brain itself (in which case they are called primary malignant tumor) or they could originate from elsewhere in the body and spread to the brain (in which case they are called secondary malignant tumor) [22]. Glioblastoma is a primary malignant brain tumor developed from star-shaped cells, called astrocytes that support nerve cells. Glioblastoma, is an archetypal example of a heterogeneous cancer and one of the most lethal human malignancies [21]. Intratumoral heterogeneity and redundant signaling routes likely underlie the inability of conventional and targeted therapies to achieve long-term remissions[9, 15, 16]. DNA and RNA profiles of bulk tumors have enabled genetic and transcriptional classification of glioblastomas. However, the relationships between different sources of intratumoral heterogeneity: genetic, transcriptional and functional, remain under research. Inter-patient variation and molecular diversity of neoplastic cells within individual glioblastoma has been previously described [19], showing that established glioblastoma subtype classifiers are variably expressed across individual cells within a tumor and demonstrate the potential prognostic implications of such intratumoral heterogeneity.

Deep Learning (DL) is a subfield of machine learning based on learning multiple levels of representations by making a hierarchy of features where the higher levels are defined from the lower levels. DL structure extends the traditional neural networks by adding more hidden layers to the network architecture between the input and output layers to model more complex and nonlinear relationship [15].

In this paper we use deep learning to perform an automated single cell classification using a dataset from 430 cells from five primary glioblastomas [19], and measure its performance.

The structure of this paper is organized as follows: Section II described the steps of the materials and methods, section III presents the experimental results and discussion and the conclusion and future work is given in section IV.

2 MATERIALS AND METHODS

2.1 Deep Learning

The concept of deep learning originated from artificial neural network research. A multilayer perceptron with many hidden layers is

EATIS 2020, May 13–15, 2020, Portugal

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

a good example of the models with deep architectures. Deep learning techniques have been applied to a wide variety of problems in recent years [14, 26]. In many of these applications, algorithms based on deep learning have surpassed the previous state-of-art performance. At the heart of all deep learning algorithms is the domain independent idea of using hierarchical layers of learned abstraction to efficiently accomplish high-level task. There are several theoretical frameworks for deep learning, and here we summarize the feedforward architecture used by H2O [1]. Multilayer perceptron (MLP) are feed-forward neural networks with architecture composed of the input layer, the hidden layer and the output layer. Each layer is formed from small units known as neurons. Neurons in the input layer receive the input data **X** and distribute them forward to the rest of the network. In the next layers, each neuron receives a signal, which is a weighted sum of the outputs of the nodes in the previous layer. Inside each neuron, an activation function is used to control the input. Such a network determines a non-linear mapping from an input vector to the output vector, parameterized by a set of network weights, which are referred to as the vector of weights **W**. The first step in approximating the weight parameters of the model is finding the appropriate architecture of the MLP, where the architecture is characterized by the number of hidden units, the type of activation function, as well as the number of input and output variables. The second step estimates the weight parameters using the training set. Training estimates the weight vector **W** to ensure that the output is as close to the target vector as possible. The structure of a MLP network is shown in Figure 1.

3 TOOLS

Library H2O [1] was used in order to perform the classification through deep learning. H2O deep learning is based on a multi-layer feedforward artificial neural network that is trained with stochastic gradient descent using back-propagation. A feedforward artificial neural network (ANN) model, also known as deep neural network (DNN) or multi-layer perceptron (MLP), is the most common type of Deep Neural Network and the only type that is supported natively in H2O.

In this study we trained a DNN using Anaconda Navigator v4.5.11 y Python 3.6. We stopped the training process after stabilization of the validation accuracy with equal weight for all the classes (1000 epochs). The batch size used is 20 samples. The network weights are initialized randomly, and the ADADELTA adaptive learning rate algorithm is used for weight updates with default parameters. The selected loss function is the categorical cross entropy.

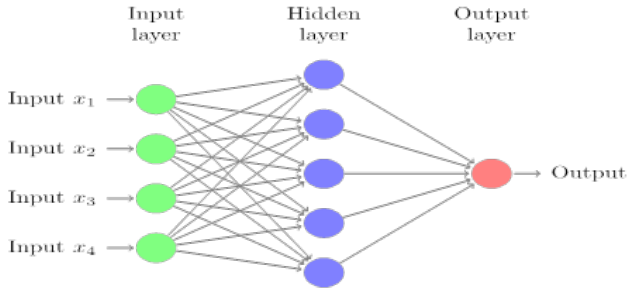
3.1 Database

The dataset consists of 430 single gliobastomas cells isolated from 5 five individual tumors [19]. The matrix data to be procesed contains 5948 rows (genes) quantified in 430 samples (columns). This database has been deposited with the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) under accession code GSE57872.

3.2 Supervised Classification of Single Cells

In order to apply the deep learning methodology for the classification of single cells a dataset for training and testing with classes

Figure 1: Structure of an architecture multilayer perceptron



was used. The dataset contains 5 classes, where each class considers all the singles cells from a primary tumor.

4 EXPERIMENTAL RESULTS AND DISCUSSION

The experimental took place using two strategies in order to build the predictive model and evaluate the accuracy of the DNN algorithm:

- Strategy I: DNN algorithm on a set of randomly selected samples, approximately 80% of the entire dataset was used for training, and approximately 20% was used as the testing set.
- Strategy II: DNN algorithm using 3-fold cross validation.

Table 1 shows the size of the DNN architecture and parameters used on the experiments to evaluate the performance of the classification. Using these parameters allowed achieve higher classification accuracy. ReLU is the non-linear activation function, epochs correspond to the numbers of passes over the training dataset, and nfolds are the number of cross-validation folds.

Variables	Parameters
input nodes	430
hidden layer	(250,250,250)
output nodes	5
activation function	ReLU
loss function	Cross-entropy
epochs	1000
nfolds	3

Table 1: Parameters of the DNN architecture

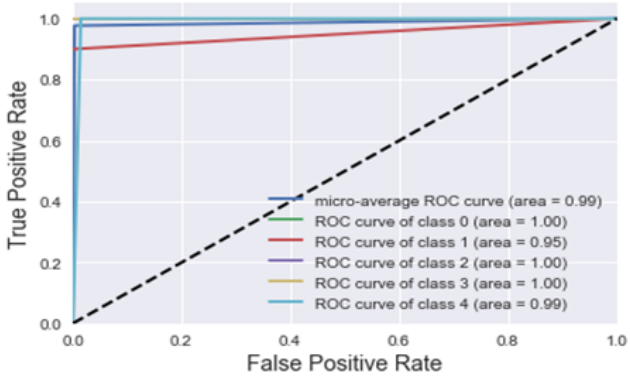
The proposed methodology was evaluated in terms of average classification rate and average area under the ROC curve (AUC) of all the 5 classes (Tumor 1 - class 0, Tumor 2 – class 1, Tumor 3 – class 2, Tumor 4 – class3, and Tumor 5 – class 4). Table 2 shows, using strategy 1, the mean square error obtained for training data and testing data, respectively.

Table 2 shows the results, using the strategy II, of the accuracy and mean square error obtained during the classification using 3-fold-cross validation. Figure 2 shows the results of the average area under the ROC curve (AUC) of all the 5 classes. We can see that

Strategy	Mode	Quantity	Value
I	Training	MSE	5.7716e-21
I	Testing	MSE	0.02
II		Accuracy	0.988
II		MSE Cross-validation	0.029

Table 2: Training record

Figure 2: Area under the ROC curve



most of the curves follows the left-hand border and then the top border of the ROC space, showing the predictive model has high precision. Additionally, the results from Table 3 show the power of this predictive model.

	Precision	Recall	F_1-score	Support
0	0.97	1.00	0.98	30
1	1.00	1.00	1.00	20
2	1.00	1.00	1.00	11
3	1.00	1.00	1.00	7
4	1.00	0.94	0.97	18

Table 3: Classification Results

The H2O deep learning framework provides the option of computing and returning the relative variable importance scores in descending order of importance. Table 4 shows the results (first 10 variables) of variable importance score for the strategy II, we can observe that DNN is able to identify which predictor variables are the most important to make these predictions.

We can see in the top of the Table 4 the predictor variable AQP4. Recently, accumulated evidence has pointed to AQP4 as a key gene that could play a critical role in glioma development [13].

5 CONCLUSIONS

In this paper we employed a deep neural network using gene expression features to classify 5 types of glioblastomas. The results show high accuracy discrimination between the classes.

Variable	Relative Importance
AQP4	1.00
CADPS	0.98
SGK1	0.97
AXL	0.97
DPP6	0.96
NUDT4	0.95
CRB1	0.95
IGDCC4	0.95
JAG1	0.95
ARHGAP26	0.94

Table 4: Variable importance scores

Using machine learning we were able to identify the most important gene - AQP4, which has been recognized by researchers as playing a significant role in glioma malignancies.

The good results achieved using this computational approach could be employed to evaluate the relationships between different sources of intratumoral heterogeneity in glioblastomas.

6 ACKNOWLEDGEMENT

This work is supported by the Hewlett Packard Enterprise Data Science Institute, University of Houston

REFERENCES

[1] Spencer Aiello, Cliff Click, Hank Roark, Ludi Rehak, and J Lanford. 2016. Machine Learning with Python and H20. *H2O. ai Inc* (2016).
[2] Philipp Angerer, Lukas Simon, Sophie Tritschler, F Alexander Wolf, David Fischer, and Fabian J Theis. 2017. Single cells make big data: New challenges and opportunities in transcriptomics. *Current Opinion in Systems Biology* 4 (2017), 85–91.
[3] Philippe L Bedard, Aaron R Hansen, Mark J Ratain, and Lillian L Siu. 2013. Tumour heterogeneity in the clinic. *Nature* 501, 7467 (2013), 355–364.
[4] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. 2015. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology* 33, 2 (2015), 155.
[5] Woosung Chung, Hye Hyeon Eum, Hae-Ock Lee, Kyung-Min Lee, Han-Byoel Lee, Kyu-Tae Kim, Han Suk Ryu, Sangmin Kim, Jeong Eon Lee, Yeon Hee Park, and others. 2017. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature communications* 8 (2017), 15081.
[6] Gregory Driessens, Benjamin Beck, Amélie Caauwe, Benjamin D Simons, and Cédric Blanpain. 2012. Defining the mode of tumour growth by clonal analysis. *Nature* 488, 7412 (2012), 527.
[7] Kolja Eppert, Katsuto Takenaka, Eric R Lechman, Levi Waldron, Björn Nilsson, Peter Van Galen, Klaus H Metzeler, Armando Poeppl, Vicki Ling, Joseph Beyene, and others. 2011. Stem cell gene expression programs influence clinical outcome in human leukemia. *Nature medicine* 17, 9 (2011), 1086.
[8] Marco Gerlinger, Andrew J Rowan, Stuart Horswell, James Larkin, David Endesfelder, Eva Gronroos, Pierre Martinez, Nicholas Matthews, Aengus Stewart, Patrick Tarpey, and others. 2012. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England journal of medicine* 366, 10 (2012), 883–892.
[9] Mark R Gilbert, James J Dignam, Terri S Armstrong, Jeffrey S Wefel, Deborah T Blumenthal, Michael A Vogelbaum, Howard Colman, Arnab Chakravarti, Stephanie Pugh, Minhee Won, and others. 2014. A randomized trial of bevacizumab for newly diagnosed glioblastoma. *New England Journal of Medicine* 370, 8 (2014), 699–708.
[10] Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, and others. 2014. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 343, 6172 (2014), 776–779.
[11] Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C

Marioni, and Sarah A Teichmann. 2015. The technology and biology of single-cell RNA sequencing. *Molecular cell* 58, 4 (2015), 610–620.

[12] Monika S Kowalczyk, Itay Tirosh, Dirk Heckl, Tata Nageswara Rao, Atray Dixit, Brian J Haas, Rebekka K Schneider, Amy J Wagers, Benjamin L Ebert, and Aviv Regev. 2015. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome research* 25, 12 (2015), 1860–1872.

[13] Yu-Long Lan, Xun Wang, Jia-Cheng Lou, Xiao-Chi Ma, and Bo Zhang. 2017. The potential roles of aquaporin 4 in malignant gliomas. *Oncotarget* 8, 19 (2017), 32345.

[14] Martin Längkvist, Lars Karlsson, and Amy Loutfi. 2014. A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters* 42 (2014), 11–24.

[15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[16] David A Nathanson, Beatrice Gini, Jack Mottahedeh, Koppany Visnyei, Tomoyuki Koga, German Gomez, Ascia Eskin, Kiwook Hwang, Jun Wang, Kenta Masui, and others. 2014. Targeted therapy resistance mediated by dynamic regulation of extrachromosomal mutant EGFR DNA. *Science* 343, 6166 (2014), 72–76.

[17] Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, and others. 2011. Tumour evolution inferred by single-cell sequencing. *Nature* 472, 7341 (2011), 90.

[18] D Williams Parsons, Siân Jones, Xiaosong Zhang, Jimmy Cheng-Ho Lin, Rebecca J Leary, Philipp Angenendt, Parminder Mankoo, Hannah Carter, I-Mei Siu, Gary L Gallia, and others. 2008. An integrated genomic analysis of human glioblastoma multiforme. *Science* 321, 5897 (2008), 1807–1812.

[19] Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, and others. 2014. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344, 6190 (2014), 1396–1401.

[20] Arnout G Schepers, Hugo J Snippert, Daniel E Stange, Maaike van den Born, Johan H van Es, Marc van de Wetering, and Hans Clevers. 2012. Lineage tracing reveals Lgr5+ stem cell activity in mouse intestinal adenomas. *Science* 337, 6095 (2012), 730–735.

[21] Jayne M Stommel, Alec C Kimmelman, Haoqiang Ying, Roustem Nabioullin, Aditya H Ponugoti, Ruprecht Wiedemeyer, Alexander H Stegh, James E Bradner, Keith L Ligon, Cameron Brennan, and others. 2007. Coactivation of receptor tyrosine kinases affects the response of tumor cells to targeted therapies. *Science* 318, 5848 (2007), 287–290.

[22] Roger Stupp, Warren P Mason, Martin J Van Den Bent, Michael Weller, Barbara Fisher, Martin JB Taphoorn, Karl Belanger, Alba A Brandes, Christine Marosi, Ulrich Bogdahn, and others. 2005. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New England Journal of Medicine* 352, 10 (2005), 987–996.

[23] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* 32, 4 (2014), 381.

[24] Alexandra-Chloé Villani, Rahul Satija, Gary Reynolds, Siranush Sarkizova, Karthik Shekhar, James Fletcher, Morgane Griesbeck, Andrew Butler, Shiwei Zheng, Suzan Lazo, and others. 2017. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* 356, 6335 (2017), eaah4573.

[25] Shinichi Yachida, Siân Jones, Ivana Bozic, Tibor Antal, Rebecca Leary, Baojin Fu, Mihoko Kamiyama, Ralph H Hruban, James R Eshleman, Martin A Nowak, and others. 2010. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* 467, 7319 (2010), 1114.

[26] Dong Yu and Li Deng. 2010. Deep learning and its applications to signal and information processing [exploratory dsp]. *IEEE Signal Processing Magazine* 28, 1 (2010), 145–154.

[27] Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, and others. 2015. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347, 6226 (2015), 1138–1142.

[28] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, and others. 2017. Massively parallel digital transcriptional profiling of single cells. *Nature communications* 8 (2017), 14049.