Taylor & Francis Word Template for journal articles

Author Namea\* and A. N. Authorb

aDepartment, University, City, Country; bDepartment, University, City, Country

Provide full correspondence details here including e-mail for the \*corresponding author

Provide short biographical notes on all contributors here if the journal requires them.

Repeat the title of your article here

Astrocytoma differentiation is poorly defined, with current methods relying primarily on histology. Here we applied computational methods to the Curated Microarray Database (CuMiDa) to examine the similarities and differences between normal samples, pilocytic astrocytomas, medulloblastomas, ependymomas, and glioblastomas. We determined the relative contribution of each transcript towards astrocytoma differentiation and applied dimensionality reduction techniques to examine the separability of each tumour subtype. Further, we explored the accuracy of various statistical models for subtype prediction via microarray data.

Keywords: astrocytoma, clustering, glioma, modelling

# Heading 1: use this style for level one headings

\* The dataset I have chosen to work with this semester is the GSE50161 dataset from the paper "Expression data from human brain tumors and human normal brain" from Griesinger et. al. The original dataset can be found here: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50161. The processed data can be found here: https://www.kaggle.com/brunogrisci/brain-cancer-gene-expression-cumida.

\* For this dataset, my primary question is whether or not we can determine the subtype of astrocytoma (a form of brain cancer) using gene expression data (Supervised Learning). The binary subquestion will be to simply determine malignancy through the expression data. The second question I would like to ask is how well the phenotype correlates with the genotype (Unsupervised Learning), and what genes are primarily responsible for the phenotypic variance. For this problem, this means assessing the distance between a sample and its cluster.

\* The dataset is a wide matrix, with over 50,000 variables for just 130 samples. This poses an interesting challenge for machine learning, and I will likely have to first learn then apply regularization or dimensional reduction to prevent overfitting. Each row is the gene expression of profile of sample of human cells that has 1 of 5 labels corresponding to the grade and presence of astrocytoma (Grades I --> Grade IV or normal). Each column represents the (normalized) expression of each gene.

\* The outcome of the project is to hopefully develop a tool that can categorize the severity of an astrocytoma based on its gene expression profile.

sever

# Excuetive Summary

\* Summarize the key (This could be a bulleted list)

+ information about your data set

+ major data cleaning

+ findings from EDA

+ Model output

+ Overall conclusions

"Expression data from human brain tumors and human normal brain" from Griesinger et. al.

# Abstract

Astrocytoma differentiation is poorly defined, with current methods relying primarily on histology. Correlations between the molecular phenotype and grading are tenuous.

\* Summary of the nature, finding and meaning of your data analysis project.

\* 1 paragraph written summary of your data analysis project

# Introduction

# Data Science Methods

Broadly, data analysis will focus on dimensionality reduction and clustering of samples in order to establish boundaries between astrocytoma subtypes. This will be done using `PCA`, `tSNE`, and `UMAP`. PCA of the subtype averages will also be used to tease out genes that are the most significant contributors to astrocytoma progression.

Subsequently, classification will be performed in several different manners. Logistic regression of dimensionally reduced data can be used to determine if a patient does or does not have cancer. `KNN` or `random forest` can we used to draw visualizable decision boundaries and to determine the similarities between different subtypes (there is currently significant debate on the categorization of astrocytoma based on histology alone). Finally, a `neural network` will be used to create a rudamentary classifier.

OUTLINE:

\* astrocytoma background

\* data from which paper

\* variable discussion

\* potential use cases of these models (e.g. autodiagnosis)

\* how to improve model

Paragraph: use this for the first paragraph in a section, or to continue after an extract.

New paragraph: use this style when you need to begin a new paragraph.

Display quotations of over 40 words, or as needed.

* For bulleted lists

1. For numbered lists

Displayed equation ( )

## Heading 2: use this style for level two headings

### Heading 3: use this style for level three headings

*Heading 4: create the heading in italics*. Run the text on after a punctuation mark.

Acknowledgements, avoiding identifying any of the authors prior to peer review

1. This is a note. The style name is Footnotes, but it can also be applied to endnotes.

References: see the journal’s instructions for authors for details on style

Table 1. Type your title here. Obtain permission and include the acknowledgement required by the copyright holder if a table is being reproduced from another source.

Figure 1. Type your caption here. Obtain permission and include the acknowledgement required by the copyright holder if a figure is being reproduced from another source.