Machine Learning for Natural Sciences

Syed Imami - 2020113012
Urvish Pujara - 2020101032

Problem statement

Analysis of Omics data using OmiEmbed

OmiEmbed supports multiple tasks for omics data including dimensionality reduction, tumor type classification, multi-omics integration, demographic and clinical feature reconstruction, and survival prediction. This can further be extended to age predictions, demographic analysis such as frequency of people at various stages of cancer in all types of cancer and estimating the growth rate of tumour. By incorporating clinical features, the model aims to provide a more comprehensive analysis that takes into account factors that can affect patient outcomes. Overall, the new model aims to improve patient care by providing accurate and personalized predictions of survival outcomes.

Recent literature on the selected problem

- **OmiEmbed: A Unified Multi-Task Deep Learning Framework for Multi-Omics Data**
  by Xiaoyu Zhang, ORCID, Yuting Xing, Kai Sun and Yike Guo
  Published on:18 June 2021
  Link: https://www.mdpi.com/2072-6694/13/12/3047
  Journal: Cancers - Volume 13, Issue 12

- **Performance Comparison of Deep Learning Autoencoders for Cancer Subtype Detection Using Multi-Omics Data**, by Edian F. Franco,22 April 2021
  Published on: 22, April 2021
  Link: https://pubmed.ncbi.nlm.nih.gov/33921978/
  Journal: Cancers (Basel)

- **Representation Learning to Effectively Integrate and Interpret Omics Data** by Sara Masarone, The Alan Turing Institute Queen Mary University of London London, UK,2022
  Published in: 2022
  Link: https://openreview.net/pdf?id=FRE7FT9DDAj
  Journal: Review paper
- **Multi-omics Data Integration, Interpretation, and Its Application** Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere2 and Krishanpal Anamika
  Published in: 2020
  Link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7003173/
  Journal: Bioinform Biol Insights , v.14; 2020
- **Self-omics: A Self-supervised Learning Framework for Multi-omics Cancer Data** by Sayed Hashim, Karthik Nandakumar and Mohammad Yaqub Mohamed Bin Zayed University of Artificial Intelligence Abu Dhabi, UAE, 2022
  Published in: 2022
  Link: https://arxiv.org/abs/2210.00825
  Journal: Archived Paper

Baseline Model

The baseline model in the new paper extends the "OmiEmbed" framework to perform tumour classification, multi omics integration, dimensionality integration, predict survival outcomes for patients with a specific disease. The omics data is preprocessed by removing unwanted features from the raw data. Variational autoencoder is used to encode and decode omics data of gene expression, DNA methylation and miRNA. Downstream tasks like diagnostic task(tumour type classification, primary site prediction and disease stage), prognostic task(prediction of gender) and demographic task(survival prediction) are then performed on the encoded data.

Proposed Improvements to Baseline

The main focus of this paper is trying to improve upon the "OmiEmbed" framework by incorporating clinical features to obtain demographic analysis of patients. Using age prediction, we can provide a high-end demographic analysis of cancer patients. This includes analysis of people at different stages of cancer suffering from various types of cancer thereby providing the average age for all stages. We can provide the feature importance for all types of cancer. Tumour growth rate can be estimated so that patients can be treated as fast as possible.