Detecting Diseases through Genetic Analysis using Machine Learning

Team 13: Haveela E J Ramakuri, Shravan B Doda, Prateek Pravanjan

28th February, 2024

1 Problem Statement

Genetic mutations and variations can provide key insights into the development and prognosis of human diseases. However, interpreting the relationship between genetics and diseases is an overly complex task. This project aims to apply machine learning techniques to genetic data in order to identify patterns that can predict disease risk and outcomes. The key goals are:

1. To develop models that can accurately classify genetic samples as belonging to either healthy individuals or those with a specific disease. Multiple disease targets will be explored, including cancer, autoimmune disorders, and cardiovascular diseases.
2. To identify specific genetic variants and combinations thereof that are highly predictive of disease status. This will provide biological insights into disease mechanisms as well as help develop easy-to-interpret decision rules based on genetic markers.
3. To compare the effectiveness of different machine learning algorithms at detecting disease-associated genetic patterns.

As students, we are inspired by cutting-edge research in genetic disease prediction, though our own work may not reach such advanced levels. However, we will make our best efforts to take inspiration from recent studies while developing novel approaches of our own. In selecting datasets and evaluation metrics, we will be guided by current standards in the field. All research papers referenced will be properly cited in the final project report.

2 Description of Dataset

This project will utilize only publicly available and freely accessible datasets related to human genetics and disease status. Potential data sources include:

1. **The Cancer Genome Atlas (TCGA)**: Contains comprehensive maps of key genomic changes in 33 types of cancer. A rich multi-dimensional open access data source. [Website: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga]
2. **UK Biobank**: Contains extensive health and genetic data on over 500,000 UK residents aged 40-69. Will require application for access but provides one of the largest human genetic datasets with linkages to medical outcomes. [Website: https://www.ukbiobank.ac.uk/]

3. **DisGeNET**: One of the largest publicly available collections of genes and variants associated with human diseases. Contains data on over 20,000 genes from expert-curated sources as well as text mining scientific literature. [Website: https://www.disgenet.org/home/]
4. **NCBI Datasets**: A collection of biomedical datasets from many sources, including genetics, genomics, imaging, and more. Freely available for searching and downloading. [Website: https://www.ncbi.nlm.nih.gov/datasets/]
5. Any additional major open access databases from the National Institutes of Health, or academic collaborations.

Note: Due to time constraints, we may opt to utilize only one of the datasets mentioned above if it fulfills our requirements. We'll decide on the dataset after our initial data exploration, ensuring it meets our analysis and modeling needs.

We aim to create our disease prediction models using freely available data only. It's crucial to have a diverse and ample training dataset for effective model performance. Hence, we'll focus on optimizing the use of public data sources before considering any private datasets. Our main goal is to make the most of open data in building our models.

3 Implementation Plan

Given the 7 to 8 week timeline, this project will focus on applying a select set of machine learning techniques on the available public genetic and disease datasets. The tentative timeline is as follows:

Weeks 1-2:

- Assemble datasets from sources described above.
- Perform exploratory data analysis to understand features, distributions, correlations, etc.
- Familiarize ourselves with the ongoing research and review the recent techniques, engineering methods etc. based on key papers in the field.

Weeks 3-4:

- Preprocess the data, including handling missing values, encoding categorical variables, and scaling numerical features.
- Implement the selected machine learning algorithms using appropriate libraries.
- Train initial versions of the models (1 algorithm per person) using a portion of the dataset and evaluate their performance using suitable metrics.

Weeks 5-6:

- Fine-tune hyperparameters of the models using techniques such as grid search or random search.
- Explore feature selection techniques to identify the most relevant features for model prediction.
- Address overfitting or underfitting issues by adjusting model complexity or regularization techniques.

Weeks 7-8

- Evaluate the performance of the refined models using cross-validation or a separate validation dataset.
- Compare the performance of different models and select the best-performing one(s).
- Document the entire process, including data preprocessing steps, model development, evaluation metrics, and results.
- Prepare a final project report summarizing the methodology, findings, limitations, and future work.

Note: The above plan is tentative and subject to change. Additional tasks, such as feature engineering, may be required during the project. Moreover, documenting choices made during the project's progression, as well as submitting interim work for feedback and evaluation, may be necessary for project success and alignment with goals.

4 Team members & task allocation

Our team consists of 3 members:

- Haveela E Joycy Ramakuri
- Shravan Bharat Doda
- Prateek Pravanjan

The key project tasks will be allocated as follows:

Individual tasks:

- **Haveela E Joycy Ramakuri:**
  - Perform exploratory data analysis and assemble datasets from the TCGA.
  - Implement Algorithm 1 (to be decided), including training, fine-tuning, and comprehensive documentation.
- **Shravan Bharat Doda:**
  - Perform exploratory data analysis and assemble datasets from the UK Biobank. Additionally, handle the application process for dataset access.
  - Perform exploratory data analysis and assemble datasets from the NCBI dataset.
  - Implement Algorithm 2 (to be decided), including training, fine-tuning, and comprehensive documentation.
- **Prateek Pravanjan:**
  - Perform exploratory data analysis and assemble datasets from DisGeNET.
  - Implement Algorithm 3 (to be decided), including training, fine-tuning, and comprehensive documentation.

Shared tasks:

- Literature review & research on state-of-the-art techniques
- Final project documentation & report

Each team member will be responsible for documenting and describing their implementation of the assigned algorithm. However, decisions regarding which algorithms to implement and evaluation methodology will be made jointly through team discussions. All members will collaborate closely on the literature review, data preparation, final analysis, and project report.