Student Checklist (1A)

This form is required for ALL projects.

1. a. Student/Team Leader: Saanvi Subramanian Grade: 12

Email: vi.iyer07@gmail.com Phone: 4254719994

b. Team Member: Kaustubh Grama c. Team Member: ______

2. Title of Project: MedulloModel: Building a Novel Machine Learning Approach to Analyze Single-cell RNA

3. School: Bothell High School School Phone: 425-408-7000

(if multiple schools, list of the team leader or list all schools).

School Address: 9130 NE 180th StreetBothell, WA 98011-3398

4. Adult Sponsor: Ashmitha Rajendran Phone/Email: ashmitha@uw.edu

5. Does this project need SRC/IRB/IACUC or other pre-approval? ☒ Yes □ No Tentative start date: ______

6. Is this a continuation/progression from a previous year? □ Yes ☒ No

If Yes:

a. Attach the previous year's □ Abstract **and** □ Research Plan/Project Summary

b. Explain how this project is new and different from previous years on

□ Continuation/Research Progression Form (7)

7. This year's experimentation/data collection:

02/01/25 03/01/25

Actual Start Date: (mm/dd/yy) End Date: (mm/dd/yy)

8. Where will you conduct your experimentation? (check all that apply)

☑ Research Institution □ School □ Field ☑ Home □ Other: ______

9. Source of Data:

□ Collected self/mentor ☑ Other Describe/url: https://www.science.org/doi/10.1126/

10. List the name and address of all non-home and non-school work site(s), whether you worked there virtually or on-site:

Name Seattle Childrens Research Institute

Address: 1920 Terry Ave, Seattle, WA 98101

Phone/email (206) 884-7200

11. **Complete a Research Plan/Project Summary following the Research Plan/Project Summary instructions and attach to this form.**

12. **An abstract is required for all projects after experimentation.**

Research Plan/Project Summary Instructions

A complete Research Plan/Project Summary is required for ALL projects and must accompany Student Checklist (1A).

- All projects must have a Research Plan/Project Summary
  - a. The Research Plan is to be written prior to experimentation following the instructions below to detail the rationale, research question(s), methodology, and risk assessment of the proposed research.
  - b. If changes are made during the research, such changes can be added to the original research plan as an addendum, recognizing that some changes may require returning to the IRB or SRC for appropriate review and approvals. If no additional approvals are required, this addendum serves as a project summary to explain research that was conducted.
  - c. If no changes are made from the original research plan, no project summary is required.
    - Some studies, such as an engineering design or mathematics projects, will be less detailed in the initial project plan and will change through the course of research. If such changes occur, a project summary that explains what was done is required and can be appended to the original research plan.
    - The Research Plan/Project Summary should include the following:
      - a. **RATIONALE:** Include a brief synopsis of the background that supports your research problem and explain why this research is important and if applicable, explain any societal impact of your research.
      - b. **RESEARCH QUESTION(S), HYPOTHESIS(ES), ENGINEERING GOAL(S), EXPECTED OUTCOMES:** How is this based on the rationale described above?
      - c. Describe the following in detail:
        - **List of materials:**
        - **Procedures:** Detail all procedures and experimental design including methods for data collection, and when applicable, the source of data used. Describe only your project. Do not include work done by mentor or others.
        - **Risk and Safety:** Identify any potential risks and safety precautions needed.
        - **Data Analysis:** Describe the procedures you will use to analyze the data/results.
      - d. **BIBLIOGRAPHY:** List major references (e.g. science journal articles, books, internet sites) from your literature review. If you plan to use vertebrate animals, one of these references must be an animal care reference.

Items 1–4 below are subject-specific guidelines for additional items to be included in your research plan/project summary as applicable.

1. Human participants research:

- **a. Participants:** Describe age range, gender, racial/ethnic composition of participants. Identify vulnerable populations (minors, pregnant women, prisoners, mentally disabled or economically disadvantaged).
- **b. Recruitment:** Where will you find your participants? How will they be invited to participate?
- **c. Methods:** What will participants be asked to do? Will you use any surveys, questionnaires or tests? If yes and not your own, how did you obtain? Did it require permissions? If so, explain. What is the frequency and length of time involved for each subject?
- **d. Risk Assessment:** What are the risks or potential discomforts (physical, psychological, time involved, social, legal, etc.) to participants? How will you minimize risks? List any benefits to society or participants.
- **e. Protection of Privacy:** Will identifiable information (e.g., names, telephone numbers, birth dates, email addresses) be collected? Will data be confidential/anonymous? If anonymous, describe how the data will be collected. If not anonymous, what procedures are in place for safeguarding confidentiality? Where will data be stored? Who will have access to the data? What will you do with the data after the study?
- **f. Informed Consent Process:** Describe how you will inform participants about the purpose of the study, what they will be asked to do, that their participation is voluntary and they have the right to stop at any time.

2. Vertebrate animal research:

- a. Discuss potential ALTERNATIVES to vertebrate animal use and present justification for use of vertebrates.
- b. Explain potential impact or contribution of this research.
- c. Detail all procedures to be used, including methods used to minimize potential discomfort, distress, pain and injury to the animals and detailed chemical concentrations and drug dosages.
- d. Detail animal numbers, species, strain, sex, age, source, etc., include justification of the numbers planned.
- e. Describe housing and oversight of daily care.
- f. Discuss disposition of the animals at the end of the study.

- **Potentially hazardous biological agents research:**
  - a. Give source of the organism and describe BSL assessment process and BSL determination.
  - b. Detail safety precautions and discuss methods of disposal.

4. Hazardous chemicals, activities & devices:

- a. Describe Risk Assessment process, supervision, safety precautions and methods of disposal.
- b. Material Safety Data Sheets are not necessary to submit with paperwork.

MedulloModel: Building a Novel Machine Learning Approach to Analyze Single-cell RNA Sequencing Data to Detect Medulloblastoma Tumorigenesis

Saanvi Subramanian and Kaustubh Grama

RATIONALE:

Provide a summary of your research. Highlight why it is important/interesting and describe any social/societal impact.

Medulloblastoma is one of the most prevalent malignant pediatric brain tumors, accounting for 20% of all brain tumors and 63% of intracranial tumors. If gone undetected, medulloblastoma can have detrimental effects on the likelihood of survival, emphasizing prompt diagnosis as a crucial step in ensuring favorable outcomes. We identified Single-cell RNA sequencing because it allows for the examination of gene expression from individual cells, provides insights into different cell types, and aids in constructing cell developmental trajectories. We chose to analyze data taken from cells of the fetal cerebellum, as its development includes pathways similar to those observed in medulloblastoma tumorigenesis.

We planned to first pre-process the sequencing data to determine cell types using clustering and commonly-used single-cell methodologies. We then wanted to create supervised machine learning models trained from these healthy cells to predict cell types in new data sets, including various developmental days and medulloblastoma tumor data. Then, we planned to compare the accuracy and efficiency of different types of machine learning models using outcomes produced in a confusion matrix. This would effectively enable us to execute a quantitative analysis of model performance to determine which models are optimal for determining the presence or absence of medulloblastoma tumor formation patterns on the cellular level.

By creating novel machine-learning approaches to detect medulloblastoma tumorigenesis, we hope to contribute to developing diagnostic methods for these cancers. In advancing diagnosis, medical personnel have more time to create targeted and personalized treatment plans that have a higher likelihood of survival before the tumor has a chance to metastasize.

Briefly explain how these relate to the rationale above.

RESEARCH QUESTION:

Which machine learning algorithm is the most optimal for detecting the presence cellular developmental trajectories indicative of medulloblastoma tumorigenesis?

HYPOTHESIS:

We hypothesize that there will be discrepancies in the accuracies of various machine learning algorithms in their classification of cellular trajectories in neurodevelopmental datasets which we can analyze to determine which is most optimal for detecting those trajectories known to be implicated in medulloblastoma tumorigenesis.

ENGINEERING GOAL(S):

We aim to engineer machine learning algorithms using Seurat and RStudio that can classify cellular developmental trajectories implicated in medulloblastoma tumorigenesis using genetic markers in fetal cerebellar Single-cell RNA sequencing data.

EXPECTED OUTCOMES:

The expected outcomes of this project include the successful preprocessing and clustering of single-cell RNA sequencing data from the fetal cerebellum to accurately identify cell types and developmental trajectories. By training supervised machine learning algorithms on annotated healthy neurodevelopmental datasets, we aim to identify genetic contributors and classify cellular trajectories associated with medulloblastoma tumorigenesis. Our analysis of the machine learning models we develop will determine the most optimal algorithm for detecting cellular patterns indicative of medulloblastoma. Ultimately, we expect this project to contribute to the development of improved diagnostic tools for medulloblastoma to enable early detection and eventual targeted treatment to increase the likelihood of survival from this pediatric brain tumor.

PROCEDURE:
Explain in some detail all procedures and experimental design created by you – do not include any work done by mentors or others associated with your project. Be sure to include information about your data collection methods.

1) Complete the necessary pre-processing on the open source data
    a) Normalizing the data
    b) Feature selection
    c) Scaling
    d) Dimensionality Reduction
2) Integrating the processed sc-RNA-seq data from the multiple timepoints
3) Identify top genes for each cell type
    a) Isolate genes with the largest effect on a cell being a specific cell type
    b) Use LogOdds metric to isolate genes with the largest effect on a cell being a specific cell type
    c) Use Variance Inflation Factor to examine multicollinearity
4) Optimize gene selection
    a) Gene selection and AIC Feature Selection
5) Create random forest and generalized linear models
6) Train model with data from days 115 and 125
    a) Initial training phase is with data from 75% of Day 115 + 125 Expression Data for each of top genes
    b) Initial testing phase is with data from 25% Day 115 + 125 of Total Expression Data for each of top genes
7) Test model with data
    a) Validation and fine-tuning with data from day 110 expression data
    b) Validation with data from day 94 expression data

8) Evaluate and compare model performance

RISK AND SAFETY:

Briefly identify potential risks and the safety procedures taken to minimize those risks.

When working with open-source data analysis software, it is necessary to acknowledge their use and origin to minimize the risk of plagiarism. Since our analysis will be performed entirely on our personal devices, there are minimal safety risks associated with this project.

DATA ANALYSIS:

Briefly describe the analysis performed on the data collected.

We will be analyzing the accuracy of the machine learning algorithms by creating a confusion matrix that compiles the prediction results into a table. The metrics consist of sensitivity, specificity, positive predicted value, negative predicted value, and accuracy for each cell type. This allows us to compare the performance of each model adjacent to each other through quantitative methods. As a supplement, we will also use visual representations to illustrate the predictions using stacked bar plots and PCA and UMAP dimensionality reduction plots created with the ggplot2 and Seurat packages in RStudio.

BIBLIOGRAPHY:

The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses (PMID: 27322403; Citations: 3,306)
Stelzer G, Rosen R, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Iny Stein T, Nudel R, Lieder I, Mazor Y, Kaplan S, Dahary, D, Warshawsky D, Guan - Golan Y, Kohn A, Rappaport N, Safran M, and Lancet D

Cope, E. C., Briones, B. A., Brockett, A. T., Martinez, S., Vigneron, P. A., Opendak, M., Wang, S. S., & Gould, E. (2016). Immature Neurons and Radial Glia, But Not Astrocytes or Microglia, Are Altered in Adult Cntnap2 and Shank3 Mice, Models of Autism. eNeuro, 3(5), ENEURO.0196-16.2016. https://doi.org/10.1523/ENEURO.0196-16.2016

Pan, J., Ma, N., Yu, B. et al. Transcriptomic profiling of microglia and astrocytes throughout aging. J Neuroinflammation 17, 97 (2020). https://doi.org/10.1186/s12974-020-01774-9

Junyue Cao et al. A human cell atlas of fetal gene expression.Science370,eaba7721(2020).DOI:10.1126/science.aba7721

Linnerbauer, M., Lößlein, L., Farrenkopf, D., Vandrey, O., Tsaktanis, T., Naumann, U., & Rothhammer, V. (2022). Astrocyte-Derived Pleiotrophin Mitigates Late-Stage Autoimmune CNS Inflammation. Frontiers in immunology, 12, 800128. https://doi.org/10.3389/fimmu.2021.800128

Monteagudo, A., Feola, J., Natola, H., Ji, C., Pröschel, C., & Johnson, G. V. W. (2018). Depletion of astrocytic transglutaminase 2 improves injury outcomes. Molecular and cellular neurosciences, 92, 128–136. https://doi.org/10.1016/j.mcn.2018.06.007