Building a machine learning model to predict intestinal function based on gut bacteria

Shashwat Choudhry
Abhinav Bhushan
May 7, 2024
Illinois Institute of Technology

Project Information:

a) Project Title: Building a machine learning model to predict intestinal function based on gut bacteria

b) Applicant name: Choudhry, Shashwat

c) Applicant major: Computer Engineering

d) Applicant Year at IIT: Rising Junior

e) Faculty Mentor(s): Bhushan, Abhinav, Armour College of Engineering

f) Previous Program Experience: None

Project Description:

a) Background Information and Problem Statement

As an undergraduate student at IIT, I have been able to understand how important undergraduate contribution can be to research projects and how through initiative I can take advantage of these opportunities. I've also come to understand the critical roles emerging sciences play in solving real-world health issues. Inflammatory Bowel Diseases (IBD) such as Crohn's disease and ulcerative colitis impact millions globally, manifesting through severe disruptions in intestinal function and significant challenges in drug absorption and metabolism. Recent studies, including those by Priya et al. (2022), have begun to unravel how specific bacterial genes might influence these processes. However, there remains a substantial gap in our understanding of which bacterial species are beneficial or detrimental in the context of IBD and how they specifically alter drug response. My prior research experience has highlighted the complex intersection between various types of data and how to harness them for predictive modeling not only in my field of computer engineering but economics as well. Working closely with a Ph.D. student on machine learning algorithms at the University of Albany, I've gained critical insights into the challenges and opportunities presented by unbalanced datasets and discrete dependent variables. This work not only honed my technical skills in algorithm development but also enhanced my capacity to communicate across disciplines, furthermore providing a robust foundation for my current research(proposal) into the microbiome's role in drug metabolism within its contexts. At the moment the proposed research project is in its phase of getting ready to create and then apply a model.

b) Objectives and Scientific Hypothesis

Building on this foundation, the goal of my project is to develop an AI/ML model capable of predicting how different bacterial species affect drug metabolism and absorption in IBD patients. I hypothesize that the interaction between certain bacterial species and human genes related to drug metabolism pathways is predictive of the pharmacokinetic profiles of medications used in

IBD treatment. This hypothesis stems from (given)initial data suggesting significant microbiome influences on gene expression linked to drug absorption and metabolism.

c) Methodology and Approach

The methodology I intend to use is a blend of computational and experimental techniques, informed by my previous work on machine learning projects(various Neural Networks):

1. **Data Collection and Analysis**: Leveraging RNA sequence data and bacterial species abundance from IBD patient samples as detailed in - Priya et al. (2022).

2. **Initial Modeling**: Identifying gene expression changes relevant to drug absorption and metabolism in IBD, correlating these changes with bacterial species prevalence.

3. **Advanced Analytical Techniques**: Implementing Principal Component Analysis (PCA) to identify key bacterial components impacting gene pathways.

4. **Model Development and Testing**: Constructing and refining AI/ML models, including regression models and neural networks, to accurately predict bacterial impacts on drug absorption and metabolism.

d) Expected Contributions and Impact

The intended contributions of my research are targeted at redefining treatment paradigms in IBD. By clarifying the role of the microbiome in drug absorption and metabolism, the AI/ML model I aim to develop could personalize and optimize treatment strategies, significantly enhancing patient outcomes. This project not only aligns with my academic focus on machine learning and its applications but also offers a tangible avenue to impact healthcare in gastrointestinal diseases, marking a critical step in my journey as a researcher in the application side of machine learning

1. **Understanding the Data**: Leveraging the given data stored in the multi dimensional spreadsheet, I plan to take time to understand the data and how I can find the

2. **Implementing Regression Model(part 1)**: Based on the data and further understanding the ability of the gene expression change(s), create a model either through a preliminary clustering and association based model. This model will have the initial characteristics of supervised learning to help the eventual final model to reduce errors in its predictive nature. This initial model will be a sort a base model the future more accurate models will be based upon.

3. **Implementing ML Model(part 2)**: Implementing a Neural Network: through the previous step with the supervised learning the next step will be to implement a form of reinforced learning of our model. I plan to base it on the existing data and research of drug absorption and the correlation of its impact on the bacterial species prevalence. The secondary model will have the original underlying data structure and the basic modeling pattern from the previous step as well. Using the underlying data patterns the model will

base itself around a classification(since we we are looking for the absorption ability).This will then eventually shift over to a form of unsupervised learning in our model.

4. **Network Testing**: Considering the various types of neural networks and their individual pro and cons alongside with future scalability and their maintenance in the future. Similar research would recommend that a Dependency-based Convolutional Neural Network would be the ideal implementation for the basis of drug absorption (Liu et al., 2019). Although this is the approach used in the referenced research another proposition may yield better result for the dataset. I also plan on testing our data set on a Recurrent Neural Network based algorithm. The algorithm I design will save the output of the recurrent cells, where the first layer is fed to a neural network followed by its recurrent layer, where the previous information(in the previous time step) is remembered by a memory function. From there on forward propagation will be implemented. In this model when there is an erroneous prediction the learning rate will come in and make the needed changes, so over its recurring steps it gradually increasing towards the correct prediction(s) during the back propagation. A outright downside of this model is the learning rate: creating the tuning parameters will possibly require an extensive optimizing algorithm and reinforced learning period of the created model. However if created this will be useful not only for my current research period but for future work on this project or any in the open source field as well.

5. **Scalability(RNN Approach):** The ideal approach(s) of the Recurring Neural Networks as proposed by me would have some issues when dealing with large scales of data. Although the RNN model would face certain challenges the outcome of an effective RNN model on large/complex data sets will be substantial. Since RNNs have the capacity of capturing temporal dynamics and dependencies that other models may overlook the predictive accuracy of our model especially when its application in in drug absorption can be significantly enhanced.

6. **Scalability(Dependency Based-CNN Approach):** In the previous research I mentioned a dependency based CNN was used effectively. Taking this approach would leverage the strengths of a convolutional neural network for a structured grid data. The D-CNN would make sure to sequence the data through various methods(I.e. parameter sharing, which unlike the RNN approach would reduce the models complexity and memory requirements). An issue that may arise is the need for a large amount of labeled data, for an effective training protocol this approach needs a large and labeled dataset. The D-CNN would also have limited effectiveness for sequential data(since it requires labeled datasets to achieve any generalizable performance any varied inputs)

7. **Model Creation and Verification(over an ideal 3 semester cycle):** The verification and enhancement of the model is what is the major backbone of my(planned) contributions to Dr. Bhushan's research project. The ideal timeline is over a 3 semester Armour Research Cycle, culminating in a optimized model that ensures the predictions have a chance to be reliable and clinically relevant. Within the summer semester of this project I plan to implement a preliminary testing. The first step being creating a simple regrading and clustering based model to understand the data and its variability on the gene expression

changes(with drug metabolism variables). The first semester will focus on creating a baseline for model performance using standard metrics(I.e. Root Mean Square Error) to create a regression accuracy and silhouette score for the clustering qualities. I plan to focus on data cleaning and a preprocessing phase focused on tailoring the data for machine learning applications. A basic implementation of the regression model o predict initial outcomes based on known gene expressions and absorption qualities. Then a utilization of clustering principles to find patters or groups within the data that may correlate with specified metabolic profiles. Once this is completed I plane to initiate the the D-CNN, which will handle the structured data and test its efficacy in capturing the spatial dependencies within the microbiome data related to the drug absorption. I will also keep in mind the scalability with a focus on memory usage to ensure it is still efficient and effective. Within the next Armour Research cycle I plan to take on the Model Optimization and Reinforcement. This cycle complexly focuses on the the training of both model. The RNN and D-CNN implementation will take preliminary insights from the initial models then shift each CNN to its optimization phase. The third and final semester will focus on Final Testing and Model Validation. I plan on having an integration phase to test how well the RNN and D-CNN models preform in a combined scenario where outputs from one model can potentially enchase the inputs or predictions of the other. Then I also will create an iterative feedback for future research continuation by other students or contributors. The loop will attempt to consider real world clinical feedback to adjust to model parameters and functions. The continuous learning implementation will allow models to adapt over time to new data without a complete retraining period.Throughout the entire three semester cycle I plan on using my engineering background and attention to detail, to create extensive documentation over the entire research period. I also alongside Dr. Bhushan, plan to showcase our progress in at the Armour R&D expo and the CAURS symposium

Student Background:

My passion for engineering, programming, and my aspirations to innovate within the field of computer engineering drive my desire to contribute to the Armour R&D program. My educational and experiential background in engineering and machine learning provides a strong foundation for conducting impactful research that aligns with the objectives of the Armour R&D initiative.

I am currently a rising junior at the Illinois Institute of Technology, pursuing a Bachelor of Engineering in Electrical & Computer Engineering, with an expected graduation in 2026. My academic journey began at California High School, where I graduated in 2022. During my academic tenure, I have engaged deeply with subjects central to computer engineering, emphasizing practical and theoretical aspects of programming and technology to understand my field and its real world relations.

My research experience so far as an undergraduate research intern has been instrumental in shaping my understanding of machine learning applications and the intersection between my field and others. Working under the guidance of a Ph.D. student at the University at Albany, I delved into predictive errors and signal extraction with discrete dependent variables. This role not only enhanced my technical skills in developing machine-learning algorithms but also honed my ability to address complex data challenges, particularly in managing unbalanced datasets.

I have fortified my programming expertise through several rigorous courses, including "C for Everyone: Structured Programming" from the University of California, Santa Cruz, and a comprehensive "C Programming Boot Camp." These courses, coupled with my hands-on project work, such as the Databasing Final Project and an ECE White Paper on ASICS and FPGA, have prepared me well for the technical demands of advanced research.

My proposed contribution to Armour R&D centers on leveraging artificial intelligence and machine learning to explore the microbiome's role in drug absorption and metabolism for inflammatory bowel disease (IBD) treatment. The AI/ML model I plan to develop aims to personalize and optimize treatment strategies, thereby attempting potentially revolutionizing patient care in this domain. I am excited about the possibility of contributing to the Armour R&D program, where I can apply my technical skills and passion for machine learning to meaningful healthcare challenges. I am confident that my background, skills, and dedicated approach will enable me to make a significant contribution to the program and further my educational and professional goals.

Thank you for considering my application.

Faculty Mentoring Plan for a Three-Semester Cycle and a Summer Cycle

Project Title:
Building a machine learning model to predict intestinal function based on gut bacteria

Project Goal:
To build an AI/ML model to predict the impact of bacterial species on human intestinal function in inflammatory bowel diseases (IBD), focusing on drug absorption and metabolism.

Faculty Background:
The faculty mentor(Dr. Bhushan) specializes in using microfluidics and microfabrication to engineer tissues and develop sensors for chemical and biochemical measurements, with a focus on diseases such as inflammation and metabolic disorders.

Student Background and Contribution:
As a Computer Engineering student with prior experience in machine learning, particularly in predictive models and handling unbalanced data, the student will develop and test an AI/ML model to redefine treatment paradigms in IBD. This project aligns with both the student's academic focus and the mentor's expertise in disease understanding and sensor application.

Three-Semester Cycle

Semester 1 - 3:Model: Development Refinement and Testing
- Objective(s): Lay the groundwork for the AI/ML model and begin initial development. From there aim to refine the model on varied data sets to asses scalability and robustness. Implement enhanced algorithms to improve model accuracy, incorporating feedback from Semester 1. From there finalize model adjustments and conduct extensive testing alongside with preparing for the Armed R&D Expo and the CAURS Symposium.

- Activities:
- Meetings twice a week to discuss project progress, integrate microfluidic data insights, and align on model structure.
- Initial data collection and preprocessing, focusing on microbial genes relevant to IBD.
- Begin understanding the regression model to establish baseline predictions.
- Initial data preprocessing, focusing on microbial genes relevant to IBD.
- Once completed begin implementation of both Neural Networks, testing each for its efficiency and ideal predictive nature
- Go through a supervised learning phase(leading into a unsupervised phase with all predictions being cross checked and verified)

Evaluation Metrics
- Progress Tracking: Meetings twice a week to discuss progress made and an variable changes to take into account

-Feedback Mechanisms: Regular and structured feedback sessions to discuss challenges, successes, and next steps.

- **Performance Reviews:** End-of-semester reviews focusing on both technical achievements and personal development. Also include various presentations and times to display findings and results. The main focus being on the predictive nature of the model and optimizing it while considering scalability

Summer Cycle

Objective: Intensive Research and Development
- Activities:
- Intensive data analysis sessions to refine model predictions
- Regular updates(ideally 2 times a week) and feedback sessions to ensure continuous progress.
- Deliverables: Enhanced prototype of the AI/ML model, comprehensive report detailing summer research advancements, preparation for subsequent semester's activities.

Cross-Cycle Resources and Support
- Academic and Research Networking: Facilitated interactions with other researchers(in program) and experts in related fields,.
- Publication and Dissemination: Guidance on preparing manuscripts for peer-reviewed journals and conference presentations.

Evaluation Metrics
- Progress Tracking: Regular assessments against established milestones.
-Feedback Mechanisms: Regular and structured feedback sessions to discuss challenges, successes, and next steps.

- **Performance Reviews:** End-of-semester reviews focusing on both technical achievements personal development.

SHASHWAT CHOUDHRY

PROFILE

am an outgoing and friendly rising junior in college, passionate about all things engineering, I have a huge passion for drawing and technical engineering documentation, and for all things programming. I aspire one day to become a Computer Engineer or Software Architect

EXPERIENCE

INTERN GLOBAL SHALA – 2022-2023

During my internship, I collaborated seamlessly with an offshore team to spearhead the development of a Python-based project. Leveraging my proficiency in the language and adeptness with simple yet powerful libraries, I actively contributed to the project's success. Engaging in cross-functional communication and remote collaboration, I navigated challenges with resilience, honed my problem-solving skills, and gained valuable insights into the dynamics of a global team.

RESEARCH INTERN – 2023

As an undergraduate contributor to a research project, I learned about the applications of ML algorithms I got to work on Predictive errors and signal extraction with discrete dependent variables with a Ph.D. student at the University at Albany. Engaging in this research project alongside a PhD student has been an invaluable learning experience, especially as an undergraduate. My primary focus has been on developing and fine-tuning machine-learning algorithms to predict discrete dependent variables and navigating the challenges posed by unbalanced data. Being able to communicate and collaborate with individuals outside my field has expanded my interdisciplinary skills, enhancing my ability to contribute meaningfully to the exploration of predictive errors and signal extraction in this context.

EDUCATION

ILLINOIS INSTITUTE OF TECHNOLOGY – ELECTRICAL & COMPUTER ENGINEERING B.E. 2025(DECEMBER)

CALIFORNIA HIGH SCHOOL – HIGH SCHOOL DIPLOMA 2022 (4.4 WEIGHTED GPA)

SKILLS & CERTIFICATIONS

As a self taught programmer alongside with my coursework in college; I have been working on these courses and projects:

- C for everyone:Structured Programming & Programming Fundamentals(U.C.S.C)
- C Programming: Modular Programming and Memory Management
- Inspirit AI Deep Dives into Designing Deep Learning Systems

Project(s):

- Databasing Final Project(completed, received grade of 99%)
- ECE Final White Paper(ASICS and FPGA paper, received grade of 98%)

References:

Priya, Sambhawa, et al. "Identification of Shared and Disease-Specific Host Gene–Microbiome Associations across Human Diseases Using Multi-Omic Integration." *Nature News*, Nature Publishing Group, 16 May 2022, www.nature.com/articles/s41564-022-01121-z.

Rohani, Narjes, and Changiz Eslahchi. "Drug-Drug Interaction Predicting by Neural Network Using Integrated Similarity." *Scientific Reports*, U.S. National Library of Medicine, 20 Sept. 2019, www.ncbi.nlm.nih.gov/pmc/articles/PMC6754439/.

Zhou, Hao, et al. "Host-Microbiome Protein-Protein Interactions Capture Disease-Relevant Pathways - Genome Biology." *BioMed Central*, BioMed Central, 4 Mar. 2022, genomebiology.biomedcentral.com/articles/10.1186/s13059-022-02643-9.