Machine Learning for Biomedical Data

Fall 2025

### COURSE DESCRIPTION:

This class is a ‘hands on’ introduction to methods and tools for machine learning for biomedical research. Topics to be included will be model training and validation, regression and regularization, unsupervised learning and clustering, dimension reduction and smoothing, supervised learning and classification, neural networks, and Bayesian learning and inference. We will generally describe the history, theory, and methods for each approach, discuss appropriate situations for application, and practice and apply computer code for applying each method. The goal of this course is to establish a fundamental understanding and working knowledge of machine learning tools, with less emphasis on mathematical rigor than other courses on campus (e.g., 14:332:443 Machine Learning for Engineers, Department of Electrical and Computer Engineering, Rutgers School of Engineering). This course is also distinct in its application on biomedical data–examples come from a variety of low and high dimensional research problems, such as epidemiology, clinical trails, biomarker discovery, and -omics data analysis. Students be expected to use R and GitHub throughout this course.

### COURSE OBJECTIVES:

Students who take this course will:

1. Become familiar with the most common methods for machine learning in biomedical research
2. Understand (generally) the theory and methods behind common methods for machine learning
3. Know the appropriate time and situations to apply each machine learning method
4. Learn the principles of model training, cross-validation, and validation
5. Gain hands on experience in applying a variety of machine learning tools to biomedical data

### PREREQUISITES

Students should have completed GSND 5345Q: Fundamentals of Data Science or have equivalent experience. A working knowledge and experience with experience R or Python is required. An introductory course in statistics, biostatistics, or equivalent experience in statistical analysis is recommended for this course.

### COURSE FORMAT:

This class will be taught virtually using a synchronous remote modality. Class will occur Tuesdays and Thursdays from 12:00-1:50pm. Courses may also be recorded and available for students who need to miss classes due to personal reasons, illness, or research related needs.

This class will be taught virtually using a synchronous remote modality, although students will be provided a classroom to gather for each lecture. A co-instructor will be present in the classroom for each lecture. Class will occur two days a week from 12:00pm-1:50pm. Courses may also be recorded and made available for students who need to miss classes due to personal reasons, illness, or research related needs.

### ZOOM Link:

Zoom ID for all sessions is 956 7664 4246 (password: 063783), or use the following link: <https://rutgers.zoom.us/j/95676644246?pwd=coPLXHRNqrf8akL5k5mSl19CDIBoiW.1>. The link is also available through the course GitHub page.

### FACULTY AND STAFF:

W. Evan Johnson, Ph.D.  
Email: [w.evan.johnson@rutgers.edu](mailto:w.evan.johnson@rutgers.edu)  
Cell Phone: (801) 472-6951

### OFFICE HOURS:

**Instructor:** Dr. Johnson will be available virtually by appointment only. Email or text him any time to set up a time to meet!

**Teaching Assistant:** There is not TA for this course. However, a co-instructor will be available for in-person help and support if needed (please reach out to Dr. Johnson if this is needed).

### GitHub REPOSITORY:

The course GitHub repository is located at: <https://github.com/wevanjohnson/2025_Fall_MLBD>. This page will contain all information in this syllabus plus more. Homework assignments and other information pertinent to this course will be posted on this web site, which will be updated frequently, so you should visit it regularly.

### CANVAS:

There will also be a Canvas course page for this course. This is where you will be able to access links to past lectures, and also turn in your homework (and track your HW grades). The rest of the course materials will only be posted on GitHub.

### COURSE TEXTBOOKS:

We will use multiple text resources in this class. None are required, all are available online or can be purchased in hard-copy. Many of my materials are adapted from these resources (thanks to the authors for these):

1. *Modern Data Science with R*, 2nd edition, By Benjamin S. Baumer, Daniel T. Kaplan, Nicholas J. Horton, Chapman and Hall/CRC, 2021. <https://mdsr-book.github.io/mdsr2e/>
2. *Introduction to Data Science: Data Analysis and Prediction Algorithms with R*, 1st edition, By Rafael A. Irizarry, Chapman and Hall/CRC, 2020. <https://rafalab.github.io/dsbook/>
3. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, 2nd edition, By Hadley Wickham, Garrett Grolemund, O’Reilly, 2017 <https://r4ds.had.co.nz>
4. *Mathematical Foundations for Data Analysis*, By Jeff M. Phillips: <https://mathfordata.github.io>.
5. *The Elements of Statistical Learning*, 2nd Edition, by Trevor Hastie, Robert Tibshirani, and Jerome Friedman: <https://hastie.su.domains/ElemStatLearn/>

### EVALUATION METHODS & COURSE GRADING

#### Assessment/Evaluation:

This course is a hands-on, project-based course. You will be graded based on homework assignments/mini projects assigned each week (10% each week, total of 70% of grade) and your final project (30%; there will be no final exam). Homework assignments and mini projects will be usually assigned at the beginning of each week and will be due by Wednesday of the week after the material is covered. However, please plan to be flexible on due dates based on the material covered in class.

#### Course Grading:

Grade Scale:

|  | 90% | 85% | 80% | 75% | 70% | <70% |
| --- | --- | --- | --- | --- | --- | --- |
| Grade | A | B+ | B | C+ | C | F |

### ATTENDANCE:

This course is being taught through a synchronous remote modality through Zoom. Attendance is mandatory; lecture recordings will only be available to students with university approved absences or pre-approved special circumstances. If you are sick or have any other justified reason to miss a lecture, please reach out to Dr. Johnson in advance and you will be reasonably accommodated.

### WORKLOAD:

This is a 2.0 credit class in the Fall 2 session. In general, you should expect four hours of in class each week, and two hours outside of class for every hour in class.

### OTHER HELP:

I **strongly** encourage you to contact early me if you have difficulty with the material. This course builds on material from prior lectures, so do not fall behind! My job is to help you understand the material as well as possible, and I am flexible with meeting times.

### ACADEMIC INTEGRITY:

You are expected to have read and follow the guidelines at the university’s academic integrity website (<http://academicintegrity.rutgers.edu> ). These principles forbid plagiarism and require that every Rutgers University student to:

* Properly acknowledge and cite all use of the ideas, results, or words of others
* Properly acknowledge all contributors to a given piece of work
* Make sure that all work submitted as his or her own in a course or other academic activity isproduced without the aid of unsanctioned materials or unsanctioned collaboration
* Treat all other students in an ethical manner, respecting their integrity and right to pursue their educational goals without interference. This requires that a student neither facilitate academic dishonesty by others nor obstruct their academic progress (reproduced from: ttp://academicintegrity.rutgers.edu/academic-integrity-at-rutgers/ ).

Violations of academic integrity will be treated in accordance with university policy, and sanctions for violations may range from no credit for the assignment, to a failing course grade to (for the most severe violations) dismissal from the university.

### COURSE TOPICS AND OUTLINE (BY WEEK)

The course will be conducted two days per week, two hours each. Tentative course start date is October 14, 2025, and the end date will be December 11, 2025.

Model training and validation (week 1)

* Evaluation metrics and cross-validation
* The bootstrap
* Application: clinical trials and drug testing

Tools for machine learning (week 2)

* The caret package in R
* The h20 package and AutoML
* Application: Biomarker development

Regression and regularization (week 3)

* Multiple regression
* Generalized linear models
* Ridge regression, LASSO, ElasticNet
* Application: infectious disease epidemiology

Unsupervised learning and clustering (week 4)

* Hierarchical clustering
* K means clustering
* Visualization: boxplots, heatmaps, etc
* Application: gene expression data

Dimension reduction and smoothing (week 5)

* Singular value decomposition, principal components
* Advanced reductions methods (NMF, UMAP)
* Kernel smoothing
* Application: single cell transcriptomics

Supervised learning and classification (week 6)

* Support Vector Machines
* Regression/decision trees
* Random Forests
* Application: genome variation analysis

Neural networks (week 7-8)

* Feed-foward and recurrent networks
* Convolutional neural networks
* Deep learning
* Application: biomarker development (revisited)

Bayesian learning and inference (week 9)

* Recurrent Neural Networks
* Generative AI, Foundation Models
* Large Language Models (LLMs)
* Agentic AI and Other AI Applications
* Application: gene regulatory networks