Team A: EHR-based sepsis prediction & personalized drug recommendations

This is a recommendation letter for your program, based on their participation in an extracurricular educational and research program offered in person in July 2023, with lectures, meetings, and mentoring sessions over the span of 2 weeks. Students worked in team projects, and thus I provide first an evaluation of the team project (section 1), and then a detailed assessment of the student and their teammates, with both a detailed evaluation and quantitative metrics (section 2). Lastly, I provide a detailed description of the course and its structure, and more details on my own background at the end of the letter (section 3).

Please do not hesitate to contact me if I can provide any additional information.

Sincerely,

Manolis Kellis, Ph.D.
Professor, MIT Computer Science
Member, Broad Institute of MIT and Harvard
MIT Stata Center, 32 Vassar St. 32D-524, Cambridge, Massachusetts, 02139, USA.
Phone: +1-617-797-4022. Email: manoli@mit.edu.

1. Team Evaluation for Team A (EHR, electronic health records)

Team Project Title:

Factors that increase the mortality of sepsis and personalized medication for patients with sepsis

Team Project Abstract:

Background: Sepsis is an acute disease with high mortality and morbidity. Early and adequate drug therapy is critical when treating patients with sepsis. Individualized treatment of sepsis is crucial considering the different patient's phenotype. Therefore this paper aims to provide guidance to physicians to better individualize the treatment plan for their patients by developing a medical prediction model to predict the mortality rate of the patient and the recommended medication for the patient.

Methods: The main methods used in this paper are Support Vector Machines, Decision Trees, Multi layer Perceptions, Random Forests, and Gradient Boosting, which are used to build a mortality prediction model using the indicators of ICU patients combined with the deaths of ICU patients, and to derive the effect of medications on the deaths of the patients.

Results: The best predictive model for mortality was Extreme Gradient Boosting with an accuracy of 0.802. The best drug prediction model was also fitted by Extreme Gradient Boosting with an accuracy of 0.867.

Conclusions: The paper concluded with machine learning methods to develop a prediction model for medication use in sepsis patients and showed that there exist 17 vital signs that affect mortality, including age, respiratory rate, etc. Also Heparin, Acetaminophen, Metoprolol Tartrate are truly effective to the treatment of sepsis patients. Norepinephrine, cefepime, furosemide is related to high mortality because most uses are on severe patients. Hydromorphone is the drug which truly decreases mortality.

Keywords: Sepsis, Personalized therapy, Mortality prediction, Drug recommendation

A copy of the report, slides, and presentation can be found at these links:

- Team A's Final Report: https://www.dropbox.com/scl/fi/0fv9jdj8we1jg9vqmwm8h/MLG23_TeamA_FinalReport.pdf?rlkey=ix98w68lqbm1r5aid9lwbv5hb&dl=0
- Team A's Presentation Slides: https://www.dropbox.com/scl/fi/r62dbexbzh2fume6vudar/MLG23_TeamA_FinalSlides.pdf?rlkey=wxp29vw63hn2729hwtwagq5lz&dl=0
- Team A's Video presentation: https://youtu.be/7X4TGIiq7Jc

Summary of project accomplishments:

Team A embarked on a study of sepsis using the MIMIC IV dataset. Their first task was to extract and process 33 relevant features for training and testing purposes. They then visualized these features to distinguish between normal, sepsis, and ICU-bound sepsis patients, in order to understand the differences between these groups.They calculated correlation coefficients and p-values for each feature in relation to mortality. Incorporating machine learning and neural network techniques, they made predictions regarding patient mortality. Their observations that medications such as Heparin and Acetaminophen might lower death rates are notable and could have practical implications. The team also examined the relationships between various drugs, their combinations, and their impact on patient death. They used ML tools to predict drug recommendations, achieving high accuracy in their predictions.

Importantly, none of the students had any prior biological knowledge, yet successfully integrated biological insights with their machine learning techniques. Their decision to study sepsis, informed in part by prior work on ARDS patients, showed their ability to take on complex medical topics.

Their thorough literature review added depth to their study. They examined up to 34 different factors that could affect sepsis patients and sought potential associations between medications, patient signs, and other related factors. Their goal was clear: to develop real-time, personalized care strategies for sepsis patients.

In terms of tools, they employed various machine learning and deep learning methods to analyze the MIMIC-IV dataset. They also made extensive use of data visualization techniques, producing a range of charts and graphs to identify and present data correlations. Their use of Overleaf for their report indicates a dedication to presenting their findings in a clear, professional manner.

In conclusion, Team A has shown dedication and skill in their study of sepsis. They combined data analytics, machine learning, and biology effectively, making significant contributions to the field in a short span of time. Their efforts, especially given their diverse backgrounds, are commendable.

Major achievements:

Team A showcased commendable achievements in their research on sepsis, employing an approach that integrated machine learning, clinical data analysis, and innovative methodologies. They extracted and processed data from the MIMIC IV dataset, which they used to make predictions. They used visualization techniques to distinguish differences in feature distributions between sepsis and non-sepsis patients, used statistical analyses to determine correlations and their significance between various features and mortality outcomes. These statistical analyses, and their use of machine learning and neural network techniques, resulted in the successful prediction of patient mortality and potential drug recommendations with high accuracy. Their results into the effects of drugs like Heparin and Acetaminophen highlighted possible correlations with reduced death rates and their hierarchical cluster analysis shed light on the interplay between different patient features. Despite having no biological background, the team combined biological reasoning and computational analysis, using numerous machine learning models, both individually and in combination, including the Extreme Gradient Boosting model. The team's results were visualized using heatmaps, charts, and figures. The data pre-processing, interpretation, and the detailed analyses they conducted, their clear presentation structure, the team's rigorous and comprehensive approach helped predict sepsis outcomes.

Areas for improvement:

Team A's sepsis project also had some limitations. They did not take drug dosage into account, focusing solely on binary use vs. non-use. They also did not consider all vital factors or potential comorbidities. Their presentation also was missing some table titles, had some redundant figure annotations, and included some graphs with very small labels, and their slides sometimes lacked clear interpretations and methodological choice explanations. Their paper sometimes lacked discussions on limitations or future directions. There was an overemphasis on the importance of sepsis and a lack of clarity regarding the chosen machine learning models. Despite using a select few drugs for their study, they missed out on creating a more comprehensive drug prediction model. Future directions can include seeking deeper insights, clearer data source definitions, and a broader approach beyond sepsis.

Suggestions for future work:

Data Enhancement: For enhancing data quality, it's recommended to include more intricate drug details, specifically dosage, and to widen the range of drugs under analysis. It would be beneficial to weave in aspects like patient complications and comorbidities to gain a well-rounded understanding of the patient's profile. Moreover, the integration of various data sources, such as medical imaging combined with Electronic Health Records (EHR), can bolster the richness of the dataset.

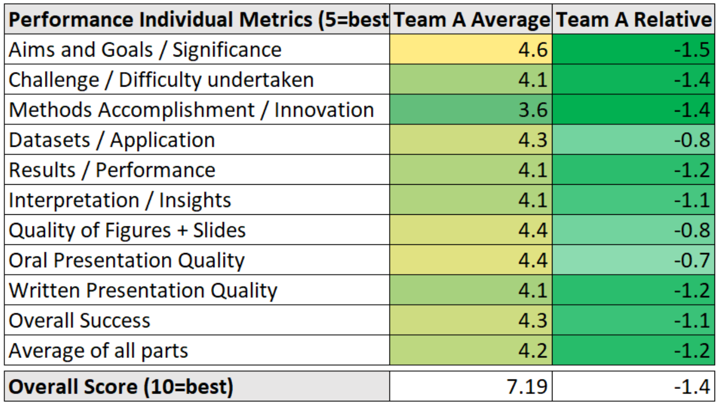
Model Refinement: When it comes to refining the model, a more in-depth exploration into chosen machine learning algorithms is suggested, highlighting the underlying rationale and specific details. Venturing into avant-garde modeling methodologies, with a particular inclination towards deep learning and models tailored for medical scenarios, could yield promising results. Furthermore, it's essential to maintain a balance between model performance and interpretability, ensuring that the model's reasoning processes remain transparent.

Figures and Interpretation. On the front of visual representation and elucidation, there's room to enhance the lucidity of graphs and figures. This can be achieved by decluttering, rectifying overlaps, or discrepancies, and honing in on salient results. To facilitate more straightforward comprehension, streamlining technical terminologies, including drug names, can prove effective.

Deepened Analysis & Report Enhancement: For a more profound analysis and report enrichment, the findings can be further underscored by infusing them with biological and clinical context. Bolstering the report with external source citations can lend more credibility, and the content can be channeled to emphasize more on the machine learning models used. The report should be meticulously structured to encompass all pertinent sections, making it comprehensive.

Practical Application & Presentation: Lastly, regarding the practical application and its presentation, strategies for external validation and ensuring a balanced dataset should be contemplated. A potential avenue worth exploring is the development of an intuitive user interface, simplifying model deployment and result interpretation. To optimize audience engagement and comprehension, refining the layout and design specifics can significantly uplift the overall user experience.

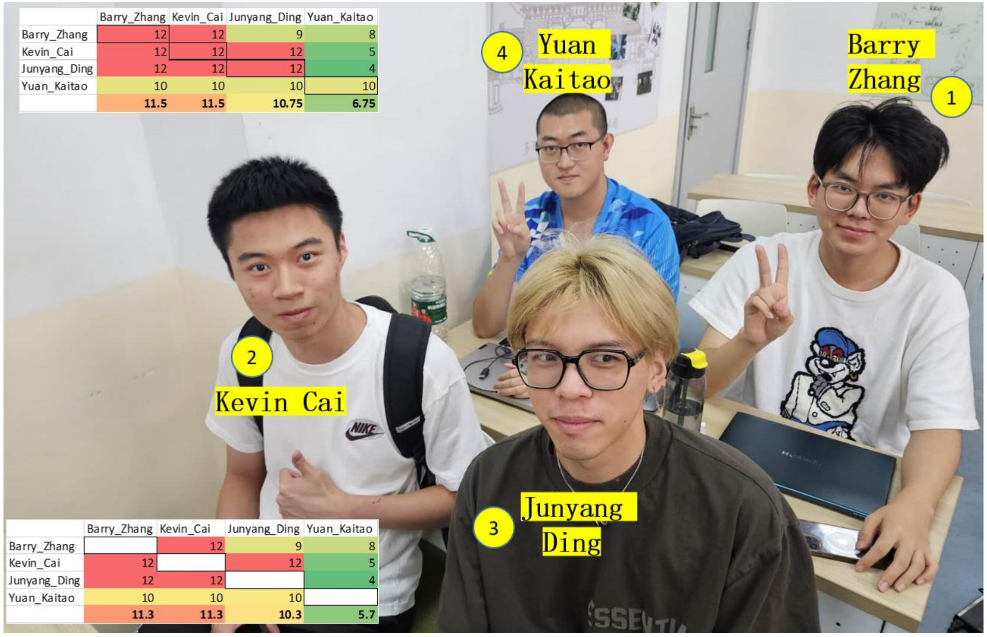
Quantitative team project evaluation:



Performance Individual Metrics (5=best	Team A Average	Team A Relative
Aims and Goals / Significance	4.6	-1.5
Challenge / Difficulty undertaken	4.1	-1.4
Methods Accomplishment / Innovation	3.6	-1.4
Datasets / Application	4.3	-0.8
Results / Performance	4.1	-1.2
Interpretation / Insights	4.1	-1.1
Quality of Figures + Slides	4.4	-0.8
Oral Presentation Quality	4.4	-0.7
Written Presentation Quality	4.1	-1.2
Overall Success	4.3	-1.1
Average of all parts	4.2	-1.2
Overall Score (10=best)	7.19	-1.4

The faculty, TAs, and all 20 students of the course provided online evaluations for each project, across 4 teams of 4-5 students each, providing a quantitative comparison for each of the following sections for each project. Team A was unfortunately the worst-performing of the four teams, about 1.4 standard deviations below the mean in overall score, and scoring particularly low in methods / innovation, given the relatively low novelty in methods. It should be noted that this was in the context of four exceptional teams that accomplished an enormous amount in the short two weeks of the course. However, these scores appropriately reflect some reservations about the overall performance of the team.

Evaluation of Team Member Barry Zhang, in the context of their prior experience and contributions:



	Barry_Zhang	Kevin_Cai	Junyang_Ding	Yuan_Kaitao
Barry_Zhang	12	12	9	8
Kevin_Cai	12	12	12	5
Junyang_Ding	12	12	12	4
Yuan_Kaitao	10	10	10	10
	11.5	**11.5**	**10.75**	**6.75**

	Barry_Zhang	Kevin_Cai	Junyang_Ding	Yuan_Kaitao
Barry_Zhang		12	9	8
Kevin_Cai	12		12	5
Junyang_Ding	12	12		4
Yuan_Kaitao	10	10	10	
	11.3	**11.3**	**10.3**	**5.7**

Barry Bairui Zhang: Score 11.3/10 (truly exceptional, joint first rank)

Background: Barry is a 19-year old third-year Computer Science student at Northeastern University at Qinhuangdao. Proficient in C++ and JAVA, he designed systems like an air ticket booking platform and a student performance management system. He's well-versed in algorithms, having explored strategies such as greedy strategy and dynamic programming. While he's explored Andrew Ng's online machine learning courses, he's currently engaged in a research project on AI-assisted diagnosis of pancreatic cancer. Barry aspires to deepen his knowledge in machine learning, aiming for an innovative final project. His long-term goal includes pursuing advanced studies in AI, potentially at institutions like the National University of Singapore or the University of Hong Kong, and then either diving into academia in China or exploring AI roles in the tech or financial sectors.

Contributions: Barry was instrumental in guiding the project, using his past experience in paper-writing and project management. Together with Kevin, he took charge of assigning tasks to the team, leveraging each member's strengths. His notable skill in time management ensured clear task timelines for all team members. On the coding front, Barry handled the MIMIC IV dataset, extracting patient features, managing missing data, and addressing outliers using box-and-whisker plots. He also coded feature density distributions for different patient groups, implemented the MLP for predictions, and applied Pearson's correlation for deeper analysis. In terms of results, Barry produced a table with 33 features, visualized feature distributions, trained the MLP model, and conducted correlation analyses. For the paper's documentation, he created 58 visuals, wrote key sections like the abstract and introduction, and ensured proper formatting and citations. Additionally, he quality-checked the paper's grammar and image formatting, offered model improvement suggestions, gave latex tutorials, and edited the final video presentation.

Evaluation: Barry exhibited a profound depth of knowledge and expertise throughout the project, evident by his high scores across various evaluations. Possessing vast experience in project participation and paper writing, Barry effectively allocated tasks to team members, leveraging their individual strengths. In the project's early stages, when the team grappled with foundational Python challenges, Barry spearheaded the data mining efforts and explored the literature to discern a fitting direction for the topic. The intricate data processing, particularly at the onset, was largely undertaken by Barry. He provided the dataset that set the stage for subsequent model predictions. His proficiency extended to the implementation of mathematical statistical methods, the MLP methods, and his significant contributions to the paper's composition, further streamlining the group's division of labor. Barry's responsibilities encompassed feature extraction, data preprocessing, and plotting. Committed to his tasks, he consistently delivered timely and high-quality work. His coding prowess was notable, with specific accomplishments including the generation of the MLP and its confusion matrix. Barry's central role in coding and pivotal contributions were instrumental to the group's overall findings and success.

TA evaluation: Barry is a diligent and accommodating teammate. Barry has multifaceted knowledge and experience in taking part in projects and writing papers, so he is not only able to assign suitable tasks for the team members well, but also in the early stage of the project he mined the data and read the literature to find a suitable direction for the topic selection. In the early stage, the data pre-processing is basically done by Barry. In the later stage, Barry implemented mathematical statistical methods and MLP methods. At the same time, he was also responsible for writing more parts of the paper to coordinate the division of labor within the group. He played an essential role in the promotion of the project.

Overall Grade and rank: Based on all these criteria, Barry Zhang was ranked #7 of all 18 students in the class, and #1 of 4 students in his team (TeamA), which was itself ranked #4/4 of the four teams, earning Barry a score of 98/100 and an A grade

Course description

Overview:

The course was offered in person over the span of 2 weeks, meeting for 2 hours daily, from Monday July 17 to Friday July 28, 2023. It was taught by Prof. Manolis Kellis, with TAs Zhu Zhang (Nanyang Technology University, Singapore, Beijing University of Technology) and Ash Zhihang Hu (Chinese University of Hong Kong, State University of New Jersey).

This course focused on applying machine learning techniques to genomic medicine, asking the students to develop an independent research project in the field of computational biology, at the interface of computer science and biology, all in the span of two weeks. The students were offered four research areas, with a sample paper in each area, and on the first day of class, four teams were formed based on student interests.

From there on, the teams were tasked with designing, planning, carrying-out, and presenting their own independent research projects, thus becoming active practitioners in a field that most students had no background in.

Project milestones:

A series of milestones enabled students to make consistent progress, thus guiding them through the typical steps of a research project, but the students were otherwise responsible for all parts of the research, with daily guidance from the professor (10am-12pm with all team receiving feedback jointly), and daily meetings with the TAs (2pm daily to help solve issues and debug).

For the first milestone (due before the first day of class), the students created a self-introductory video, template-based info sheet, and information form about their background, teaching them to present themselves in multiple formats, and providing a starting point for team formation. The students then formed teams in the first 20 mins of class, based on their levels of interest in the four papers that the TAs and Professor had suggested.

Their second milestone (due on Day 2), establishing a data/feasibility demo video, and detailed project description, literature search, paper description, and top 3 project ideas. The students were tasked to show that they can reproduce at least one figure of the assigned papers, and thus that code was available, datasets were available, and that they were able to write the necessary wrappers.

Milestone 3 (due on Day 3) was the formal project proposal, outlining specific aims both on the computational front (machine learning method development), and on the biological data front (questions and analyses planned), including the roles of each student team member based on their background and expertise.

Milestone 4 (due on Day 4 and the evening of Day 3) was a peer-review evaluation of each team's proposal by members of all other teams, with each student reviewing exactly one project (their second choice typically, thus quite aligned with their interests), and each team receiving 3-4 peer reviews. This allowed students to both receive ample and detailed feedback from their peers, and also for each student to put themselves in the professor's shoes, and think critically about their peers' projects, with their critical eye often helping them recognize improvements that they could make on their own projects. The reviews were also structured in a way that allowed members of each team to review all other proposals, and each team being reviewed by members of each other team, thus having a bi-directional flow of ideas, suggestions, and recommendations for adoption or avoidance. Students received their feedback by the evening of Day 3, and provided their responses by the morning of Day 4, providing an outline for how they will revise their proposal, aims, methods, and planned analyses.

Milestone 5 (due on Day 5) was to demonstrate continued exploratory analyses, machine learning improvements, and making substantial progress on their milestones and aims.

Milestone 6 (due on Day 6, Monday of the second week) was a Midcourse Report, wherein they evaluate initial milestones, remaining work, timelines, and start working on the introduction and methods sections of their final reports, while outlining the figures, display items, tables, and datasets of their reports.

Milestone 7 (due on Day 7 and the evening of Day 6) was a similar peer review round, with each reviewer providing additional feedback on their assigned team project, each team receiving additional feedback, and each team providing a response and revised directions if needed.

Milestone 8 (due on Day 8) didn't have any specific deliverable, except more results, results, results, providing a chance to the professor and TAs to provide scientific feedback to the students about their methods, analyses, results, and more. The students worked on analyzing data, focusing entirely on producing new 'science' based on their code, applied to their data, and interpreted by members of their teams.

Milestone 9 (due on Day 9) was to produce nearly-final figures, visualizations, insights, tables, heatmaps, and graphs that provide insights on the data and problem at hand, providing a chance to receive feedback on every aspect of their proposal, but also on the written, visual, and oral presentations.

Milestone 10 (due on Day 10, the final day of class) was final slides, report, and video presentations, with a final chance to receive feedback on every aspect of the work. This was coupled with extensive peer evaluations and feedback from all students in the class, including: (a) self-evaluations within each team, for their own project, but also the contributions of all other team members; and (b) peer-evaluations from every student to all other teams, providing detailed scoring and feedback on every aspect of the project.

Milestone 11 (due the following Friday) was giving every team a chance to finalize their results and address any feedback received by the professor, TAs, and students, given how hard the students had worked to form teams, gather biomedical data, develop methods, and analyze their datasets. This led to continued progress for some teams, but relatively little

These milestones allowed the students to gain expertise across all aspects of the research process, across inception, planning, execution, and presentation, including method development, data gathering, analysis, interpretation, visualization, oral presentation, and written presentation.

About the professor:

Manolis Kellis is a Professor of Computer Science at MIT, a member of the Broad Institute of MIT and Harvard, a member of the Computer Science and Artificial Intelligence Lab at MIT, and head of the MIT Computational Biology Group (compbio.mit.edu). His research is in the areas of disease genetics, epigenomics, gene circuitry, non-coding RNAs, comparative genomics, and phylogenomics. He has helped direct several large-scale genomics projects, including the Roadmap Epigenomics project, the ENCODE project, the Roadmap Epigenomics Project, the Genotype Tissue-Expression (GTEx) project, and comparative genomics projects in mammals, flies, and yeast. He received the US Presidential Early Career Award in Science and Engineering (PECASE) by US President Barack Obama, the Mendel Medal for Outstand Achievements in Science, the NSF CAREER award, the Alfred P. Sloan Fellowship, the Technology Review TR35 recognition, the AIT Niki Award, and the Sprowls award for the best Ph.D. thesis in computer science at MIT. He has authored over 275 journal publications, which have been cited more than 156,000 times. He lived in Greece and France before moving to the US, and he studied and conducted research at MIT, the Xerox Palo Alto Research Center, and the Cold Spring Harbor Lab. For more info, see: http://compbio.mit.edu/.