

1 Nurturing diversity and inclusion in AI in Biomedicine through a

2 virtual summer program for high school students

3

4 Tomiko Oskotsky^{1,2*}, Ruchika Bajaj^{3*}, Jillian Burchard^{4*}, Taylor Cavazos^{5*}, Ina Chen^{5*}, Will
5 Connell^{6*}, Stephanie Eanoff^{2,7*}, Tianna Grant^{5*}, Ishan Kanungo^{8*}, Karla Lindquist^{9*}, Douglas
6 Myers-Turnbull^{10,11*}, Zun Zar Chi Naing^{12,13,14,15*}, Alice Tang^{2,8,16*}, Bianca Vora^{3*}, Jon Wang^{8*},
7 Isha Karim^{17,18*}, Claire Swadling^{18,19*}, Janice Yang^{18,20*}, AI4ALL Student Cohort 2020**¹⁸, Marina
8 Sirota^{1,2}

9

10 ¹Department of Pediatrics, UCSF, San Francisco, CA, USA.

11 ²Bakar Computational Health Sciences Institute, UCSF, San Francisco, CA, USA.

12 ³Department of Bioengineering and Therapeutic Sciences, UCSF, San Francisco, CA, USA.

13 ⁴Cognitive Science and Computer Science Programs, UCLA, Los Angeles, CA, USA.

14 ⁵Program in Biological and Medical Informatics, UCSF, San Francisco, CA, USA.

15 ⁶Pharmaceutical Sciences and Pharmacogenomics, UCSF, San Francisco, CA, USA.

16 ⁷Berkeley Institute for Data Science, UC Berkeley, Berkeley, CA, USA.

17 ⁸School of Medicine, UCSF, San Francisco, CA, USA.

18 ⁹Department of Epidemiology & Biostatistics, UCSF, San Francisco, California, USA.

19 ¹⁰Institute for Neurodegenerative Diseases, UCSF, San Francisco, CA, USA.

20 ¹¹Quantitative Biosciences Consortium, UCSF, San Francisco, CA, USA.

21 ¹²QBI COVID-19 Research Group (QCRG), San Francisco, CA, USA.

22 ¹³Quantitative Biosciences Institute (QBI), UCSF, San Francisco, CA, USA.

23 ¹⁴J. David Gladstone Institutes, San Francisco, CA, USA.

24 ¹⁵Department of Cellular and Molecular Pharmacology, UCSF, San Francisco, CA, USA.

25 ¹⁶Department of Bioengineering, UC Berkeley, Berkeley, CA, USA.

26 ¹⁷Saint Francis High School, Mountain View, CA, USA.

27 ¹⁸AI4ALL, Oakland, CA, USA.

28 ¹⁹Canton High School, Canton, MI, USA.

29 ²⁰Department of Computer Science, Stanford University, Stanford, CA, USA.

30 * Denotes equal contribution

31 ** A full list of members appears in the Author Contributions

32

33

34 Abstract

35 Artificial Intelligence (AI) has the power to improve our lives through a wide variety of
36 applications, many of which fall into the healthcare space; however, a lack of diversity is
37 contributing to flawed systems that perpetuate gender and racial biases, and limit how broadly
38 AI can help people. The UCSF AI4ALL program was established in 2019 to address this issue
39 by promoting diversity and inclusion in AI. The program targets high school students from
40 underrepresented backgrounds in AI and gives them a chance to learn about AI with a focus on
41 biomedicine. In 2020, the UCSF AI4ALL three-week program was held entirely online due to the
42 COVID-19 pandemic. Thus students participated virtually to gain experience with AI, interact
43 with diverse role models in AI, and learn about advancing health through AI. Specifically, they
44 attended lectures in coding and AI, received an in-depth research experience through hands-on
45 projects exploring COVID-19, and engaged in mentoring and personal development sessions
46 with faculty, researchers, industry professionals, and undergraduate and graduate students,
47 many of whom were women and from underrepresented racial and ethnic backgrounds. At the
48 conclusion of the program, the students presented the results of their research projects at our
49 final symposium. Comparison of pre- and post-program survey responses from students
50 demonstrated that after the program, significantly more students were familiar with how to work
51 with data and to evaluate and apply machine learning algorithms. There was also a nominally
52 significant increase in the students' knowing people in AI from historically underrepresented

53 groups, feeling confident in discussing AI, and being aware of careers in AI. We found that we
54 were able to engage young students in AI via our online training program and nurture greater
55 inclusion in AI.

56

57 Introduction

58 Artificial Intelligence (AI) has the power to improve our lives through a wide variety of
59 applications. Just a few examples of how AI is being used to enrich our lives include search
60 engines, autonomous vehicles, and facial-recognition, route-planning, and ride-hailing programs
61 (1). The applications of AI to the biomedical, translational, and clinical realms are diverse
62 ranging from discovering biomarkers and repurposing therapeutics, to improving disease
63 diagnosis and automating surgery (2). Moreover, AI can help realize the promise of
64 personalized medicine, a healthcare approach that aims to tailor medical decisions and
65 treatments to individuals based on their intrinsic (e.g., genomic, age, sex) and extrinsic (e.g.,
66 diet, environmental exposures) factors (3).

67

68 Yet a lack of diversity can adversely affect how broadly AI will help people (4). For instance, if
69 machine learning (ML) algorithms to diagnose skin cancer lesions were trained on data that
70 largely represent fair-skinned populations, then the algorithms, no matter how advanced, would
71 not perform as well on images of lesions in skin of darker color (5). We need diversity not just in
72 the data we use in AI but also in the people working and leading in the field of AI. Currently, AI
73 professors are mostly male (>80%), and among AI researchers, only 15% at Facebook and
74 10% at Google are female (6). Moreover, black workers represent 4% of the workforce at
75 Facebook and Microsoft, and only 2.5% of Google's entire workforce (6).

76

77 The UCSF AI4ALL program, established in 2019 and co-directed by Marina Sirota, PhD and
78 Tomiko Oskotsky, MD, strives to promote greater diversity and inclusion in the field of AI in
79 biomedicine, and to inspire tomorrow's leaders to think about and know AI and to use AI
80 ethically. UCSF AI4ALL recruits high school students from backgrounds underrepresented in AI,
81 including females and students from minority racial and ethnic backgrounds, as well as students
82 from low income families and those who are the first in their families to go to college. Through
83 this tuition-free three-week summer training program, students gain experience with AI with a
84 focus on applications to biomedicine, interact and work with a diverse set of role models in AI,
85 including women and people of Black or African American background and Hispanic or Latino
86 background, and learn about how AI can advance health. They receive broad exposure to AI
87 topics through faculty lectures, and gain in-depth research experience through hands-on
88 projects. Mentoring and career/personal development sessions with faculty, researchers,
89 industry professionals and undergraduate and graduate students further enable personal growth
90 and an opportunity to explore career interests at the intersection of Computer Science and
91 Biomedicine. Due to the COVID-19 pandemic, the UCSF AI4ALL program held in 2020 shifted
92 from an in-person, commuter program to a synchronized, online one with all the student
93 research projects focusing on leveraging AI to advance our knowledge and understanding of
94 COVID-19. Here, we provide an overview of the 2020 UCSF AI4ALL virtual summer program,
95 share details about the research projects our students engaged in, and discuss the results of
96 our program.

97 Methods

98 We reviewed all 89 complete applications that were submitted to our program during the
99 application period in March 2020, and assessed each candidate holistically prior to offering
100 acceptances into our program.

101

102 The program itself was a three week program, which was held virtually in 2020. The first week
103 focused on teaching the high school participants basic programming in python and introductory
104 topics in machine learning and the second and third weeks focused on research projects, which
105 in 2020 were all leveraging AI to COVID-19. Google CoLab notebooks were used as a simple
106 means to share code with students and run Python. Each morning began with time for students
107 to ask questions as well as participate in our ice-breaker activities. The program had daily guest
108 lectures by diverse faculty from UCSF focused on application of AI in biomedicine and covering
109 a wide range of topics from clinical data analysis, to diagnostic and therapeutic strategies
110 leveraging molecular measurements. There were also panels composed of UCSF AI4ALL
111 student alumni, undergraduate students, graduate students, and professionals from private
112 companies with backgrounds in AI within biomedicine and other disciplines. Our panelists, many
113 of whom were women and people from diverse racial and ethnic backgrounds, spoke with our
114 students and shared insights about their work and their journeys. Each week, our Alumni TAs
115 led Community Building Session engaging the class of students in fun, bonding exercises. We
116 also held a personal growth session to develop the students' communications skills. The end of
117 the program symposium included student presentations on their research projects as well as a
118 Keynote talk on AI in Biomedicine. A copy of the 2020 program schedule is available

119 https://ai4all.ucsf.edu/assets/2020_UCSF_AI4ALL_Program_Schedule.pdf

120
121 Students were asked to complete a survey at the beginning (Pre-) and at the conclusion (Post-)
122 of our program. Mann Whitney U (MWU) test with continuity correction was used to compare
123 Pre- to Post- survey responses (since surveys were anonymous, we could not compare these
124 using tests designed for paired data), and to compare 2019 to 2020 Pre- and Post- survey
125 responses. Bonferroni corrections were employed, and a significance threshold of 0.05 was
126 applied to the results.

127

128 Results

129 Of the 89 high school students who submitted applications to our program and the 38 applicants
130 we accepted into the program, 29 enrolled in and completed the program.
131 All 29 students were females who were rising sophomores (21%), juniors (45%) or seniors
132 (34%) in high school. Most of the students were from California (79%), although several were
133 from other states. The racial backgrounds of the students included Asian inclusive of those from
134 the Indian subcontinent and Philippines (79%), Native Hawaiian or Other Pacific
135 Islander/Original Peoples (3%), and Hispanic or Latino (7%), and 14% declined to state.
136 Twenty-one percent will be first generation college students. (**Table 1**).
137

Table 1. Demographic characteristics of accepted students in the 2020 UCSF AI4ALL Summer program

Characteristic	# students (total accepted: 29)	% students
Gender		
She/Her	29	100%
He/Him	0	0%
They/Them	0	0%
Race *more than one category may be checked		
Asian (including Indian subcontinent and Philippines)	23	79%
Black or African American	0	0%
Native Hawaiian or Other Pacific Islander (Original Peoples)	1	3%
Hispanic or Latino (including Spain)	2	7%
White (including Middle Eastern)	0	0%
Decline To State	4	14%
Grade Level Next Year		
Senior / 12th grade student	10	34%
Junior / 11th grade student	13	45%
Sophomore / 10th grade student	6	21%
Freshman / 9th grade student	0	0%
Qualify for Free Lunch at School		
Yes	4	14%
No	25	86%
1st Gen College Student		
Yes	6	21%
No	23	79%
Home State		
California	23	79%
Other	6	21%

139 1st week: Lessons in Python and Machine Learning

140 In the first week of the program, students spent the afternoons learning about machine learning
141 concepts and programming in Python. We had seven UCSF graduate student instructors and
142 teaching assistants (TA) to help with teaching during the first week. iPython notebooks with the
143 in-class exercises were shared the evening before the class, to give students an opportunity to
144 practice on their own before the solutions were reviewed in class.

145

146 Python Workshops

147 Students covered the basics of programming, data management, and data visualization in the
148 first two days to prepare to code in Python language and work with data within a Google CoLab
149 environment in preparation of their projects. Topics covered include programming basics (data
150 types, logic, loops, functions), data structures, common Python packages, plotting with
151 matplotlib, and using sklearn. During the lesson, students were placed in breakout rooms with
152 teaching assistants to review coding exercises and practice programming activities together.

153

154 Lessons on Machine Learning

155 Topics covered in ML include Data and Bias, Clustering, Classification, Naive Bayes,
156 Regression, and Neural Networks. To facilitate remote instruction, we employed a reverse
157 classroom paradigm, in which the instructors produced a 15-20 minute lecture video to be
158 watched before each classroom session. The general structure of live sessions include 15
159 minutes of topic review first, then the rest of the session covering either conceptual activities or
160 reviewing and practicing ML exercises on CoLab notebooks. For activities, students were
161 placed into breakout rooms with a teaching assistant. Since students come from various
162 backgrounds of ML and programming familiarity, collaboration within the breakout rooms were
163 encouraged.

164

165 The instruction was carried out in an inverse classroom setting where the participants could ask
166 the instructors and TAs questions after having watched the lectures.

167 **2nd and 3rd weeks: Research Projects and Presentations**

168 **Applying AI to COVID-19**

169 The students were assigned to one of five groups based on their preference, each working on a
170 research project that applied AI to the characterization, classification, or prediction of COVID-19
171 leveraging different types of biomedical data - gene expression, proteomics, imaging and
172 clinical. Each project team was led by a UCSF graduate student, medical student, and/or
173 postdoctoral scholar and co-led by an alumni TA.

174

175 On the last day of the program, students shared findings from their group's research project
176 during our Final Symposium. Each presentation was approximately 15 to 20 minutes in length,
177 with time for questions, and each student presented a portion of their group's work. The event
178 was attended virtually by over 100 people, including faculty, graduate students and postdoctoral
179 scholars from UCSF and other institutions, program participants and their invited family
180 members. A videorecording of the Final Symposium, including our Keynote Speaker's talk and
181 the students' presentations is available, <https://youtu.be/uImjiHI7MDw>.

182

183 **Project 1: AI for Global Health - AI and COVID-19 Time Series**

184 **Diagnosis Data**

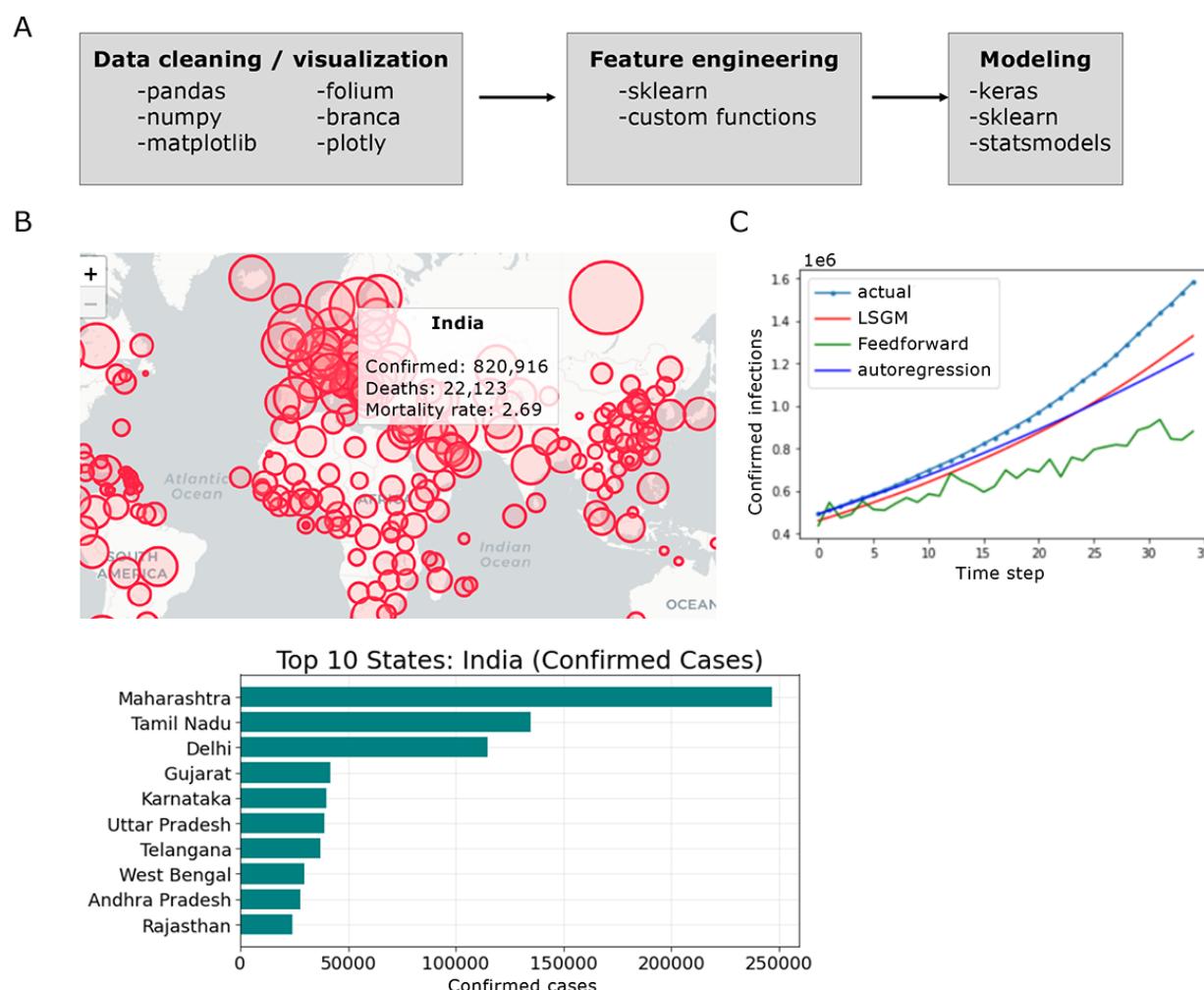
185 Students learned how to develop machine learning algorithms with utility for lower middle
186 income country (LMIC) settings. Their objective was to develop an algorithm that can predict the

187 number of cases in a given country. Students used publicly available daily time series data
188 describing confirmed COVID-19 infections and deaths per country and states across the world
189 (over 266 regions) aggregated from the Johns Hopkins Center for System Sciences
190 downloaded on July 1, 2020 (<https://github.com/CSSEGISandData>)(7). Each student then
191 manually pre-processed the dataset to a format in which they could conduct exploratory data
192 analysis. The educational approach was to allow students to have first-hand experience in
193 discovering the optimal way to plot and analyze various features of the data they were working
194 with by experimenting with different visualization libraries and troubleshooting together real-time
195 through video conferencing. Recognizing the diversity in time series trends between countries
196 during exploratory data analysis, students chose to narrow the scope of the problem to focus on
197 a specific country, selecting India due to its large number of cases and disparity in public health
198 services. Students then did a literature review to understand the public health issues in India
199 and how to design an algorithm that may actually provide utility to key stakeholders in the
200 region.

201
202 Visualizing the trend of confirmed infections in India, they decided to develop a forecasting
203 algorithm that can aid in identifying how many resources a given country or state will need.
204 Students were then presented with high-level information on several ML techniques used for
205 time series data analysis, such as autoregression (8), Holt-Winters (9) exponential smoothing,
206 and neural networks (10). Following a group debrief, students were allowed to select modeling
207 approaches that interested them. Afterwards, they trained, developed, and tested three different
208 algorithms: autoregression, feed-forward neural network, and Long Short-Term Memory (LSTM)
209 recurrent neural network (**Fig 1a,b**).

210
211 Students first started with a simple ML technique for time series data, known as autoregression.
212 Afterwards, they decided to see whether it was possible to leverage data from other countries
213 that may be useful in the same prediction. They developed a feedforward neural network

214 algorithm that leveraged data from 266 other countries/regions and predicted the most recent 5
215 days of COVID-19 cases in India. Students discovered that this model performed worse than
216 the simple ML technique (Mean Average Percentage Error: 82.35% for feedforward neural
217 network vs. 8.23% from autoregression) (**Fig 1c**). The students became interested in trying
218 LSTM recurrent neural networks, due to their unique ability to model time-series data better than
219 feed-forward networks. They trained the model to predict the next 5 days of data from the most
220 recent 15 data points, and found it performed slightly better than the feed-forward, but not as
221 well as the simple ML technique (Mean Average Percent Error: 10.08% for LSTM) (**Fig 1c**).



222
223 **Fig 1. AI for Global Health. A.** Schematic of major ML skills explored with packages/utilities
224 used for instruction. **B.** Examples of data visualizations created by students. **C.** Model
225 predictions compared to actual India COVID-19 data. Mean average percent error was 8.23%,
226 10.08%, 82.35% for autoregression, LSTM, and feedforward neural networks respectively.

227

228 **Project 2: AI and Proteomics - COVID-19 Protein-Protein Interactions**

229 **(PPI) predictions**

230 Students learned to implement supervised learning techniques to predict host protein
231 interactors, given primary amino acid sequences of various viral proteins, and to test if proteins
232 of similar sequences would interact with the same host proteins. The dataset was curated from
233 two publicly available sources - 1) a host-pathogen protein-protein interaction (PPI) data in
234 HEK293T cells for HIV (11), HCV (12), HPV (13), Ebola (14), Dengue (15), and Zika (15), which
235 contains sequence information on virus proteins with corresponding human protein information
236 and their MiST scores, i.e. their interaction confidence scores (16) and 2) human proteome fasta
237 files containing one protein sequence per gene (16). Mostly, project time was spent covering
238 data processing, support vector machines (SVMs), and deep learning using Python. The group
239 put these concepts in practice through hands-on work with their individual project; the six
240 students chose one of six pathogens (HIV, HCV, HPV, Ebola, Dengue, and Zika) to work on
241 individually. First, students built a dataframe containing their chosen virus-human interactions
242 and split the data frame into a positive dataset with host-pathogen interactions with MiST score
243 ≥ 0.75 and a negative dataset with proteins of their chosen virus and a randomized human
244 proteins, resulting in 248 positive interactions and 496 negative interactions for Dengue, 89
245 positive and 140 negative interactions for HCV, 704 positive and 1400 negative interactions for
246 Zika, 93 positive and 186 negative interactions for HIV. The students then applied a global
247 sequence alignment using a Biopython package (17) to check the pattern of the sequence
248 alignments between positive and negative interactions. To predict their virus-human protein
249 interactions, students combined the negative and positive datasets and created a SVM. The
250 students coded separately on personal Jupyter notebooks but shared code through
251 CoLaboratory notebooks and collaborated through project time discussion, screen-sharing, and
252 Slack messaging.

253

254 In the first week, the group began by reading three papers on PPI predictive techniques and
255 dissected the various merits of each paper's methodology before deciding to pursue sequence
256 alignment to demonstrate homology between their protein family and another potentially related
257 member. Next, the students accessed the PPI dataset of host-pathogen PPI data containing
258 virus bait protein, corresponding human prey protein and gene name from PubMed for their
259 chosen pathogen. In their first dataset, they mainly organized the bait and prey sequences and
260 corresponding MiST score in a comprehensive and cohesive format. Each student chose a
261 different virus of six options. Together, they collaboratively processed the primary PPI dataset
262 by isolating their virus' bait and protein sequence to build their virus-protein dataframe. To close
263 off the week, the instructors introduced the second training dataset consisting of each of the six
264 pathogens' protein ID and sequences. The students downloaded and utilized fasta files from
265 UniProt that contained the protein ID and sequences for HIV, HCV, HPV Ebola, Dengue, Zika
266 and spent the remainder of project work time understanding the relationship between prey and
267 protein sequences.

268

269 To start the second week, students learned about different sequence alignment algorithms.
270 First, the students split each pair of interacting virus bait and human prey depending on the
271 MiST score into positive (MiST ≥ 0.75) and negative (MiST < 0.75) datasets. After splitting the
272 dataset, the instructors guided the students in constructing a data processing pipeline prior to
273 building their predictive model. To add features of the data, the group utilized a global pairwise
274 alignment algorithm from Biopython (17) to add the sequence alignment scores for each bait
275 and prey pair to the positive and negative dataframes. Guided by their instructors, the group
276 deliberated and decided on features that may serve as potent predictive variables including bait
277 protein length, amino acid counts, and the atomic weight of the bait protein. Finally, the students
278 visualized the distribution of alignment scores for the positive and negative data and evaluated
279 the association.

280

281 In the final week, the group began creating and testing their machine learning models. The
282 students combined the positive and negative dataframes and visualized their data through a
283 scatter plot. Next, the group implemented SVMs and collaboratively built their classifier.
284 Students selected an 80% and 20% split for their training and testing data, respectively. Each
285 student first trained their model using their individual virus data. Then, they trained the model
286 using all their virus data to predict the interaction between each SARS-CoV-2 protein and each
287 human protein from the first PPI dataframe they built. The students finetuned the algorithmic
288 parameters, to improve the model's performance. To visualize the algorithm's optimal
289 performance, each student built a confusion matrix for the SVM predicting virus-human protein
290 interaction (**Fig 2a-e**) and extracted feature importance in a bar plot (**Fig 2f**). Additionally,
291 students were guided by their instructors to build a convolutional neural network (CNN) for their
292 individual pathogen.

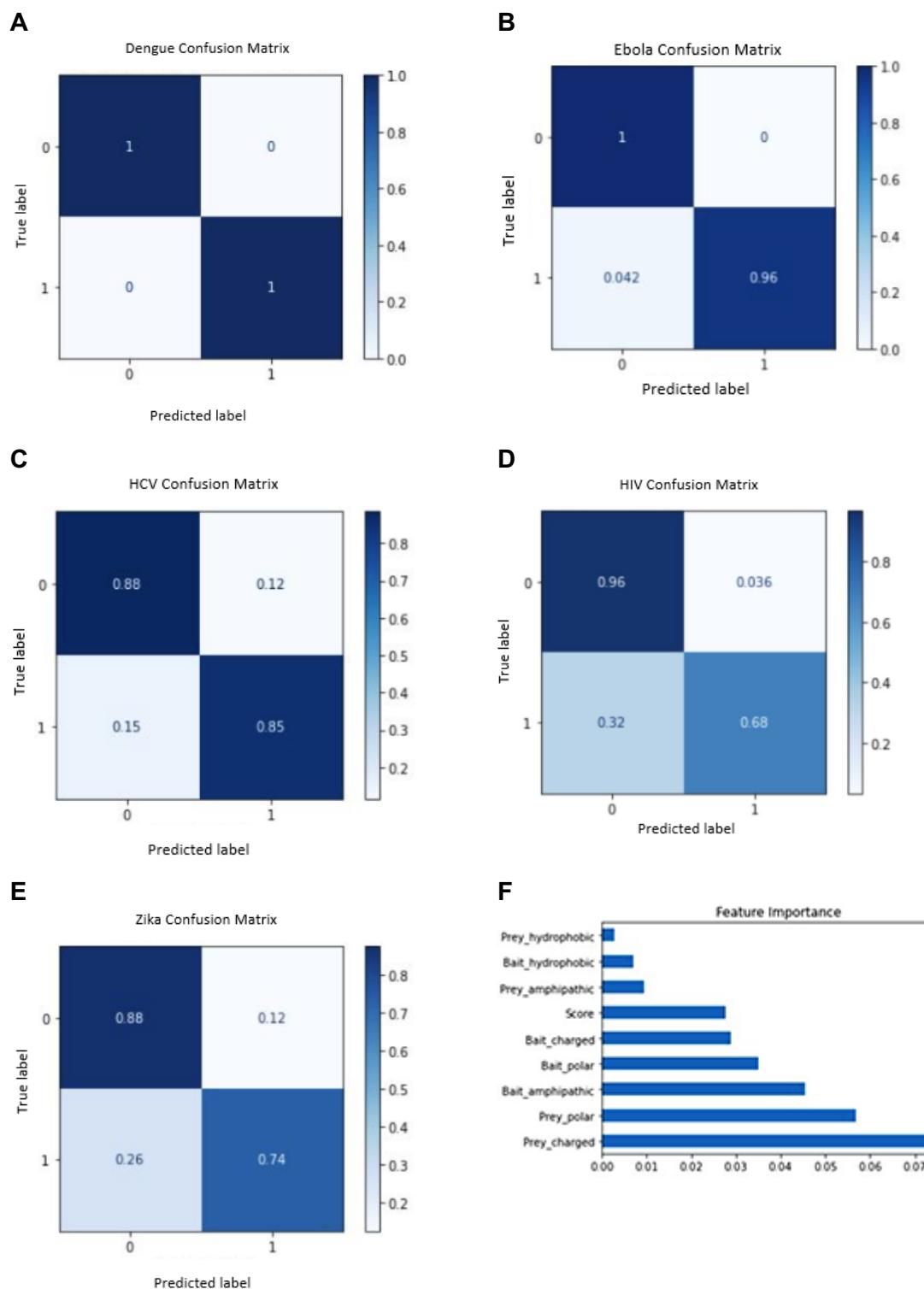
293

294 Finally, the students spent the remainder of the last week investigating the implications of their
295 research. The instructors presented ways that this type of data can be leveraged for drug
296 discovery and repurposing whereupon the students discussed the implications of their
297 experimental results on the COVID-19 pandemic.

298

299 Code availability, data, slides, and figures can be found [here](#).

300



301 **Fig 2. AI and Proteomics: COVID-19 Protein-Protein Interactions (PPI) Predictions. A-E:**
302 **Confusion matrices visualizing the individual SVM models' performances.** Evaluated the
303 true and predicted virus-human protein interactions for Dengue, Ebola, HCV, HIV, and Zika
304 viruses respectively. **F: Bar plot of feature importance extracted from Ebola virus trained**

305 **SVM on SARS-CoV-2 virus data.** The prey proteins with a net positive or negative charge had
306 the highest feature importance.

307

308 **Project 3: AI for Imaging Data - Computer Vision for COVID-19 Chest**

309 **X-Ray Classification**

310 Students learned to implement machine learning models that can classify COVID-19 cases in
311 chest x-ray images. The dataset was curated from publicly available chest x-ray images
312 datasets (18,19) and contained 438 images from patients diagnosed with COVID-19 as well as
313 438 images from patients without pathologic findings, labeled as 'no finding' (**Fig 3a-b**). The
314 instructors introduced concepts in data processing, computer vision, and deep learning using
315 Python. The students applied these concepts in a hands-on project where they first visualized
316 and evaluated the dataset then trained a convolutional neural network (CNN) to identify COVID-
317 19 cases. The students coded separately on personal CoLab notebooks but collaborated and
318 debugged together to the extent possible through online video discussions and screen-sharing.

319

320 The group began by examining the dataset through exploratory data analysis. They
321 experimented with different data visualization approaches and packages to plot the distribution
322 of attributes in the dataset. Using the visualizations, the students discovered potential biases in
323 the dataset such as skewed gender distribution and different x-ray views for the COVID-19
324 images versus the 'no finding' images. The students also learned to use dataframe
325 manipulations to survey subsets of the datasets. Together, they identified image features, such
326 as lung region opacity and body outlines, that could be useful or potentially problematic for the
327 classification task.

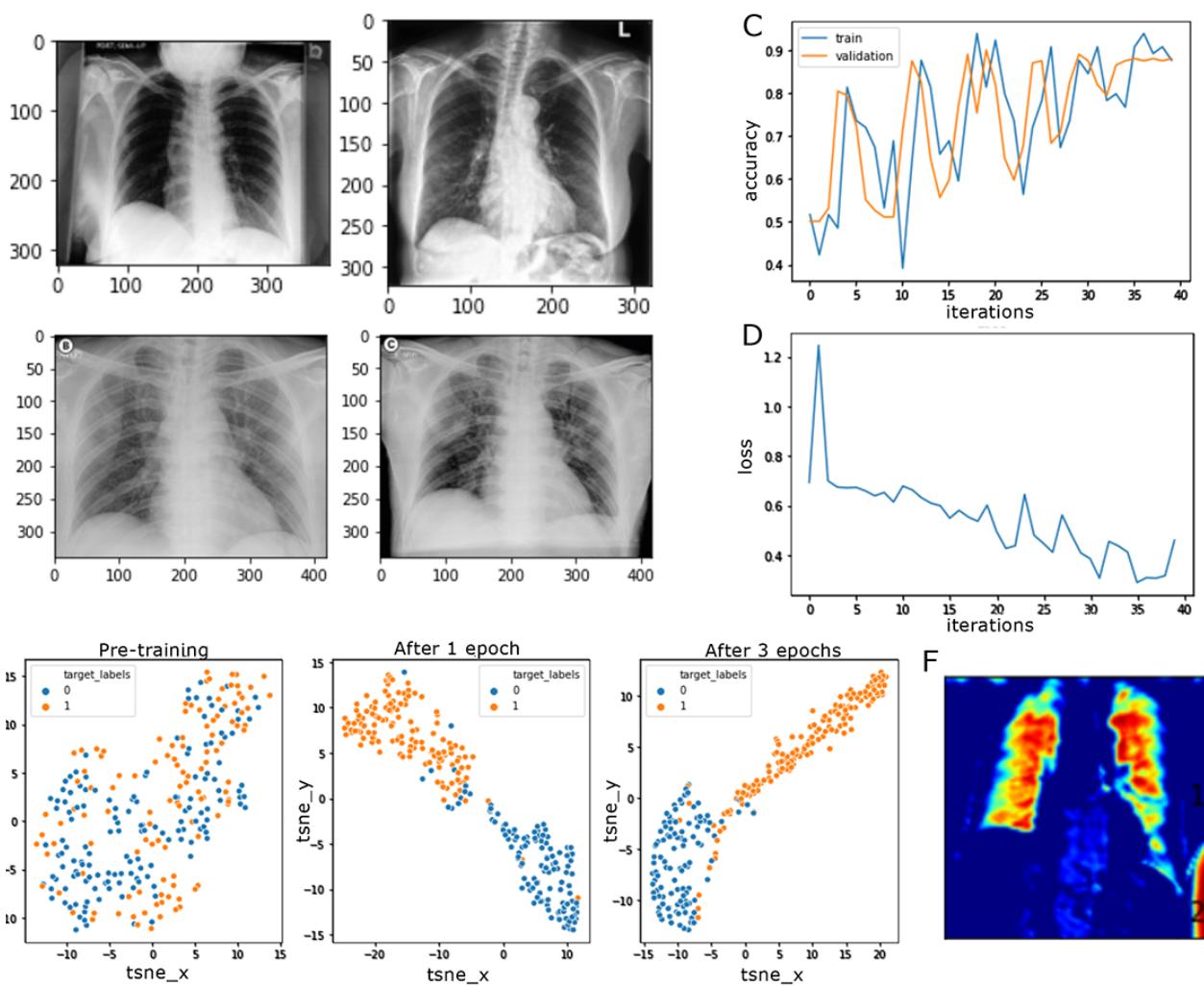
328

329 Next, the students developed fully-connected and convolutional neural networks (CNN) using

330 PyTorch to perform binary classification. The instructors provided a starting code framework that
331 guided the students in constructing a typical data processing and machine learning pipeline.
332 The students completed the code by implementing missing core sections. They tracked model
333 training speed as well as accuracy and loss curves to gain insight into the model training
334 process (**Fig 3c-d**). The students also calculated metrics, such as F1 scores, and visualized
335 latent space features to evaluate the model (**Fig 3e**). By discussing these analyses, the
336 students identified areas where the model is performing poorly (e.g. incorrectly classifying
337 COVID-19 images) and formulated hypotheses for potential reasons, such as the large
338 variability in the COVID-19 images.

339

340 In the final week, the students branched out to work on follow-up ideas. For this half of the
341 project, the students were largely self-driven to explore ideas or questions that interested them.
342 The instructors helped the students work through code issues and brainstormed solutions. The
343 follow-up ideas included iteratively improving CNN models by tuning hyperparameters and
344 training other types of classifiers (SVM, regression) using features extracted by a pre-trained
345 DenseNet model. One particular challenge the students tackled was model robustness. Using
346 gradient class activation maps to provide interpretation of model decisions, they had observed
347 that the CNN models were detecting regions outside of the lung (**Fig 3f**). This became a focus
348 for model improvement. During breaks, the instructors also briefly introduced interesting papers
349 and recent results in AI vision research. The group discussed challenges of interpretable
350 decision-making and model brittleness, which complemented the tasks they were working on.



362 **Project 4: Latent Variable Modeling of COVID-19 Metagenome**

363 **Transcriptomic Profiles**

364 Students applied dimensionality reduction techniques to investigate natural stratification of
365 SARS-CoV2+ patient subgroups based on host transcriptomic response and viral coinfection
366 status. Previous analyses of metagenomic sequencing data from upper airway samples of 238
367 patients revealed a diminished innate immune response in patients positive for SARS-CoV2
368 through differential expression analysis, gene set enrichment analysis, and *in silico* estimation of
369 cell type proportions (22). We hypothesized decomposition of metagenomic next generation
370 RNA sequencing (mNGS) to reveal separable clusters of patient subgroups. In our approach,
371 we applied both supervised and unsupervised decomposition methods to analyze structural
372 patterns inherent to the data.

373

374 First, the students accessed this publicly available data and performed exploratory analysis on
375 the study cohort of 94 patients who tested positive for SARS-CoV2 by gold standard clinical
376 PCR. Mick et. al quantified the abundance of 15,900 host genes and 275 viruses by RNA
377 sequencing and reference based alignment. Students applied principal component analysis
378 (PCA) to each dataset and qualitatively inspected emergence of clusters based on covariates
379 such as SARS-CoV2 viral load, gender, and age. In this unsupervised analysis the students
380 experienced how sample outliers can skew variance and cause inflation of PCA components. In
381 order to account for possible non-linear structure within the data, the students also trained and
382 tested autoencoders on both the viral coinfection dataset and the gene expression dataset.

383

384 Evaluation of viral coinfection embeddings led the students to hypothesize a correlation
385 between binary coinfection status and SARS-CoV2 viral load. The students reasoned that
386 patients with additional viral infection(s) may be generally more susceptible to SARS-CoV2
387 replication due to alteration of immune response dynamics under coinfection conditions.
388 Patients were stratified into 2 groups: those with measurable alternative viral load, and those

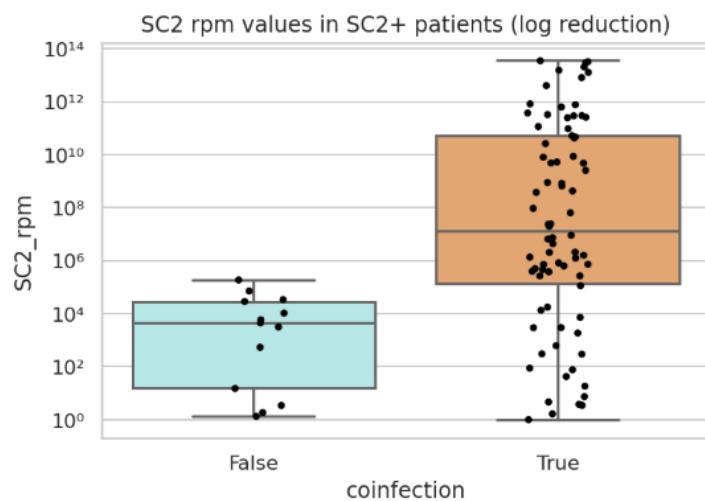
389 with sole viral metagenome alignment to SARS-CoV2. Welch's t-test was applied to the average
390 reads-per-million (rPM) of SARS-CoV2 between the two groups, and a significant difference
391 below an error tolerance of 0.05 was found ($p=1.947*10^{-4}$) (**Fig 4a**).

392

393 The students followed up this result with an analysis for confounding variables. In group
394 discussion, we posited that samples with a large SARS-CoV2 viral load cause a higher
395 frequency of alignment errors to evolutionarily similar genomes. To test this, the students
396 inspected correlation coefficients between SARS-CoV2 and alternative, coinfection viral read
397 abundance. The most correlated viruses were all coronaviruses, lending evidence to the
398 hypothesis that higher SARS-CoV2 viral load results in a higher proportion of reference
399 alignment errors to genomically similar viruses (**Fig 4b**). This experience underscored the
400 adage, "correlation does not imply causation," lending a valuable lesson to the analysis of
401 SARS-CoV2+ patient subgroups.

402

403 **A**

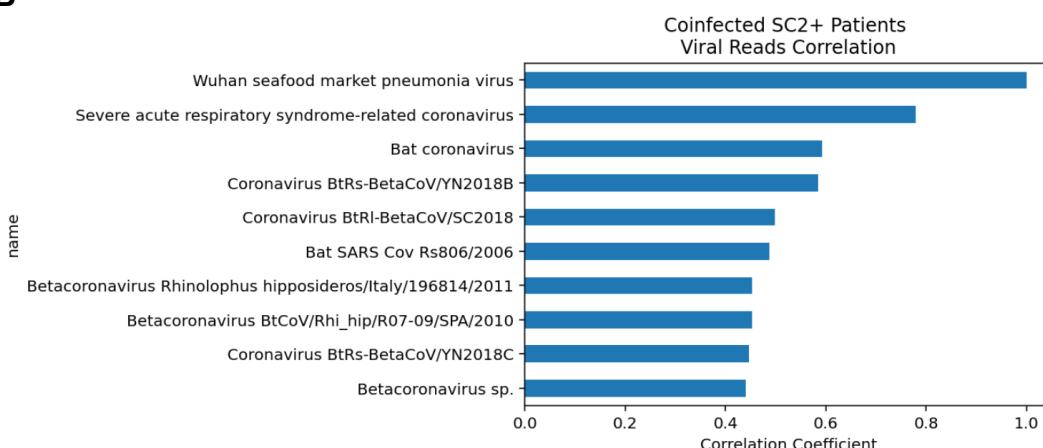


404
405
406

407

408

B



409

410 **Fig 4. Latent Variable Modeling of COVID-19 Metagenome Transcriptomic Profiles.**

411 **A. Boxplot of SARS-CoV2 rPM by coinfection status.** Reads-per-million of SARS-CoV2

412 stratified by patients with alternative (orange) and no (blue) detectable viruses.

413 **B. Correlation coefficients between reads of SARS-CoV2 and coinfected viruses.** Highly

414 correlated viruses are coronaviruses.

415

416 **Project 5: PredictCOVID - AI and Electronic Medical Record (EMR)**

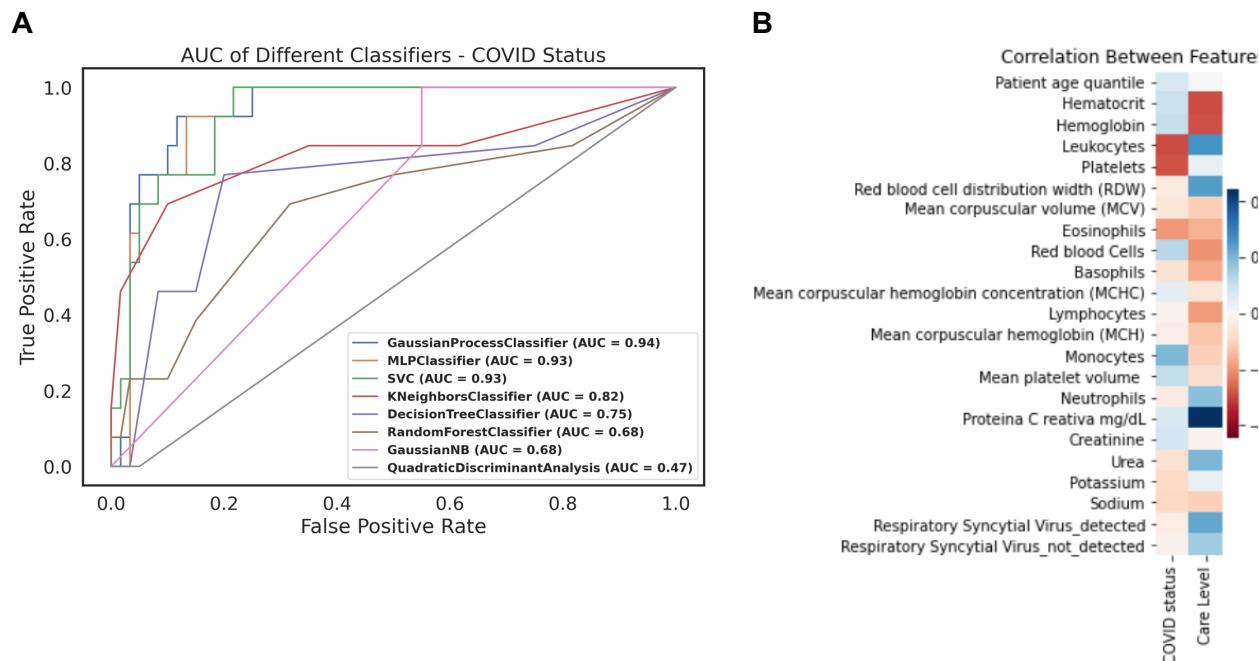
417 **data**

418 Using real world data from (<https://www.kaggle.com/einsteindata4u/covid19>), students learned
419 how to apply AI to (A) predict whether a patient is COVID positive or negative and (B) predict
420 the severity of the COVID infection (i.e. admission into the general ward, semi-intensive care
421 unit, or intensive care unit). The dataset included 5,644 patients as well as COVID-19 test
422 results, patient age quantile, hospital admission ward, and various laboratory results from blood
423 tests, urine tests, and pathogen tests. Students were introduced to the benefits and drawbacks
424 of publicly available data, such as sources of bias and the need for intensive data preprocessing
425 before the data can be utilized. Students took different approaches to data cleaning and
426 imputation of missing values and evaluated the performance of machine learning models on
427 varied input data, including evaluating metrics of accuracy, area under the curve (AUC), and
428 distribution of false negatives and positives (**Fig 5a**).

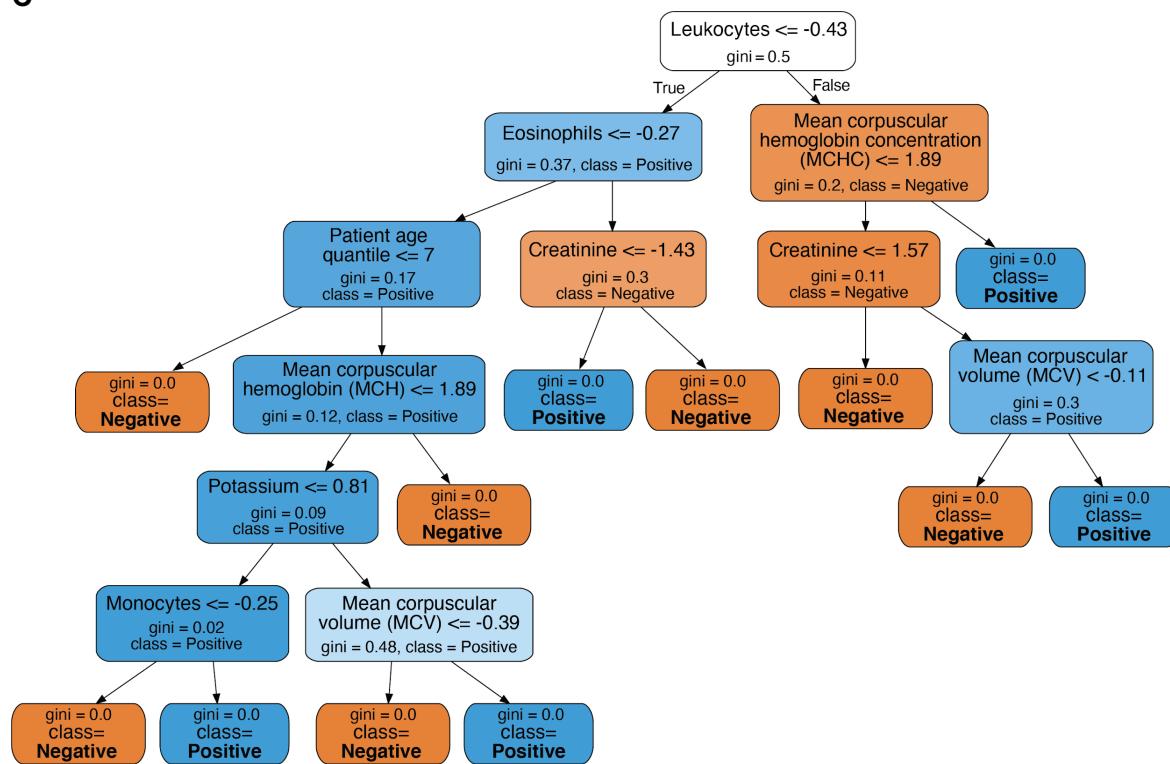
429

430 Additionally, students were asked to go above and beyond to apply their findings to translational
431 applications. For example, students were asked to critically evaluate the cost of false negatives
432 (spreading COVID-19, not receiving treatment on time, worse outcomes) and false positives
433 (waste of limited resources) in respect to patients and outcomes, and applying this evaluation to
434 the decision of a model. Students were also asked to perform covariate analyses to determine
435 feature importance and apply back to their understanding of clinical relevance and application
436 (**Fig 5b,c**). One finding that the group reported was that leukocytes were heavily negatively
437 correlated with COVID test results (**Fig 5d**). Lastly, the group summarized their findings and
438 recommendations for future plans to the entire group as well as the limitations and biases in the
439 data (i.e. single location, limited follow-up, missing data).

440

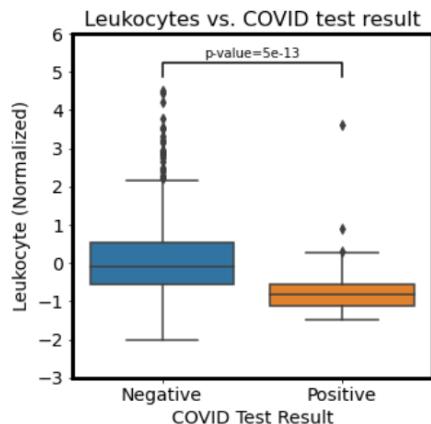


C



441

D



442

443 **Fig 5. PredictCOVID: AI for deriving insights from healthcare data. A. Area Under the**
 444 **Curves (AUCs) of different machine learning classifiers for predicting COVID status from**
 445 **patient data. AUC was one of the metrics used by the students to evaluate the performance of**
 446 **their machine learning models. B. Decision tree classifier for COVID status. Features used in**
 447 **decision making in each level of the tree are shown, with COVID status (positive/negative) at**
 448 **the leaves of the tree. C. Heatmap of the correlation between clinical features and COVID**
 449 **status or Care Level. Blue represents positive correlation while red represents negative**

450 correlation. **D. Boxplot of normalized leukocyte laboratory values and COVID status.**

451 Leukocyte laboratory values negatively correlated with COVID test results.

452

453 **Pre- and Post- Surveys**

454 Analysis of the survey data revealed significant shifts in some of the students' responses from

455 the Pre- to Post- survey. Specifically, at the end of the program, there were significantly more

456 students who reported that they know how to clean data before using it in machine learning

457 algorithms (MWU test, adjusted p-value <0.001), and know how to evaluate and apply machine

458 learning algorithms (MWU test, adjusted p-values <0.001) (**Fig 6**). More students also reported

459 knowing people in AI who are people of color (MWU test, unadjusted p-value = 0.014, adjusted

460 p-value = 0.285) and women (MWU test, unadjusted p-value = 0.044, adjusted p-value = 0.877),

461 feeling confident in questioning the media about AI (MWU test, unadjusted p-value = 0.015,

462 adjusted p-value = 0.297), and knowing about careers that use AI (MWU test, unadjusted p-

463 value = 0.037, adjusted p-value = 0.743); however, these increases were only of nominal

464 significance. Survey responses of students in the 2020 virtual program and those of students in

465 the 2019 commuter program for questions asked in both years were found not to differ

466 significantly (Mann Whitney U test, unadjusted and adjusted p-values > 0.05) (**Supplementary**

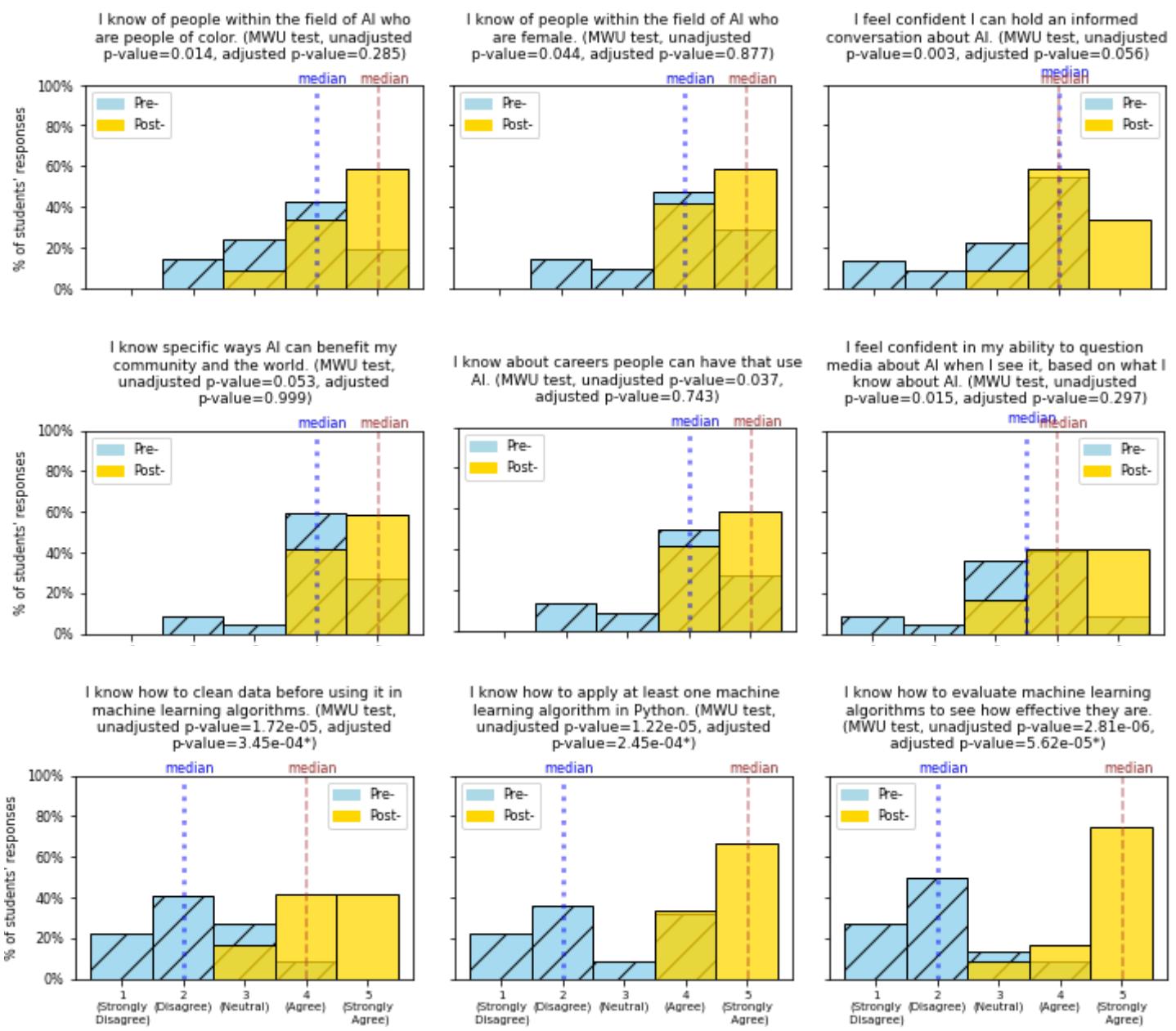
467 **Table 1**). Students of the 2020 virtual were also no less likely to recommend the AI4ALL

468 program to peers than the students who attended the 2019 in-person program (Mann Whitney U

469 test, unadjusted p-value = 0.044, adjusted p-value = 0.872) (**Supplementary Figure 1**).

470

471



472 **Fig 6. Pre-post survey results.** Histograms of students' responses to pre-program and post-
 473 program survey questions, with median values. Mann Whitney U tests were performed to
 474 compare pre- and post-survey responses, with adjusted p-values below significance threshold
 475 of 0.05 given an asterisk (*).
 476
 477 Moreover, the graduate students, post-doctoral scholar and alumni TA's who worked with our
 478 students on their research projects shared their observations that the students:

479 ○ Grew in their abilities to think critically, form hypotheses, and design executable
480 experiments
481 ○ Learned how to develop a variety of different ML models
482 ○ Fortified their collaborative skills and technical proficiency in Python
483 ○ Found exploratory data analysis to be valuable as it challenged the participants to view
484 scientific inquiry in an open-ended way that deviated from traditional classroom
485 experiences
486 ○ Learned how to analyze the results and how they relate to the original research questions
487 ○ Experienced the immediate relevancy of AI approaches to current problems in the
488 COVID-19 pandemic
489 ○ Demonstrated their understanding and interpretation of not only AI but also the
490 application of AI to medicine, public health, and clinical decision making

491 **Discussion**

492 Diverse representation is needed not just in the data for Artificial Intelligence (AI), but also in the
493 people working and leading in the field of AI. Since 2019, UCSF AI4ALL has engaged students
494 from backgrounds historically underrepresented in AI in order to promote greater diversity and
495 inclusion in this field. In 2020, through a variety of interactive virtual real-time sessions and
496 experiences held during a three-week period, our program allowed students to interact with a
497 diverse set of role models in AI and learn about how AI can be used to advance health.
498 Furthermore, students gained experience in coding, working with data, and AI by participating in
499 one of our meaningful hands-on research projects that applied AI to understanding, classifying,
500 or predicting COVID-19.

501
502 Students' survey responses demonstrated their feeling significantly more familiar with working
503 with data and evaluating and applying machine learning algorithms at the end of our 2020 virtual

504 program. There was also a nominally significant increase in the students' knowing people in AI
505 who are from historically underrepresented groups, their confidence in discussing AI, and their
506 awareness of careers in AI. While the format of the 2020 program differed from 2019, with the
507 2020 program taking place online instead of in-person due to the pandemic, students' survey
508 responses from both years were comparable.

509

510 Despite the success of our virtual training program, there were some limitations to having a
511 program take place entirely online, including the lack of in person interactions and the need for
512 reliable internet connection. Nevertheless, the ability to engage young students in AI and the
513 opportunity to contribute to diverse representation in this field make holding our program in any
514 format worthwhile.

515

516 We have learned that it is possible to deliver virtually an AI curriculum to young high school
517 students that provides them with an engaging and impactful experience. Through our virtual
518 program, we were able to connect with students from around the country and involve teaching
519 assistants and faculty from outside the Bay Area and from other institutions. We were also able
520 to give students who are located far from AI training programs a chance to become involved
521 bringing the goal of increasing diversity in AI a little closer to reality.

522

523 Author Contributions

524 TO and MS designed and co-directed the program, performed analysis of program survey data,
525 outlined and wrote the manuscript. JW, IK, and JB led and described Project 1. ZN and IK led
526 and described Project 2. IC, TG, and JY led and described Project 3. WC, RB, and CS led and
527 described Project 4. AT and BV led and described Project 5. JB, TC, WC, SE, TG, KL, AT, and

528 DMT developed curriculum materials and led instructional sessions. All authors discussed
529 results, provided critical feedback and contributed to the final manuscript.

530

531 **AI4ALL Student Cohort 2020**

532 Arhana Aatresh¹⁸, Githika Annapureddy¹⁸, Ami Baid¹⁸, Qixin Dai¹⁸, Esha Gohil¹⁸, Yomn
533 Hammad¹⁸, Ishika Hazra¹⁸, Vienna Huang¹⁸, Valerie Kwek¹⁸, Marissa Lee¹⁸, Angela Liu¹⁸, Neha
534 Naveen¹⁸, Ava Paikeday¹⁸, Sonica Prakash¹⁸, Srihita Ramini¹⁸, Yamuna Rao¹⁸, Neeharika
535 Ravi¹⁸, Mihika Rayan¹⁸, Brenda Samano¹⁸, Jessica Sanchez¹⁸, Anooshkha Shetty¹⁸, Arely Sun¹⁸,
536 Riya Tadi¹⁸, Monica Trinh¹⁸, Gianna Yan¹⁸, Joyce Yang¹⁸, Caroline Yoon¹⁸, Fiona Zhang¹⁸, and
537 Hannah Zhuang¹⁸.

538 Corresponding author: Correspondence to Marina Sirota (marina.sirota@ucsf.edu)

539

540 **Acknowledgements**

541 The authors would like to acknowledge everyone in the AI4ALL organization, including Tess
542 Posner, Olga Russakovsky, Fei Fei Li, Valerie Allen, Beth McBride, Jonathan Garcia, and
543 Thalea Torres, and our colleagues in the Bakar Computational Health Sciences Institute,
544 including Angela Jackson and Sharat Israni, for their encouragement and support. We would
545 also like to acknowledge Aeshah Al-Nagdawi, Gaia Andreoletti, Matthew Hancock, Andrew Jan,
546 Chakrapani Kalyanaraman, Idit Kosti, Brian Le, Hunter Mills, Boris Oskotsky, Ben Rubin, Wren
547 Saylor, Gundolf Schenk, Victoria Turner, Rohit Vashisht, and Katharine Yu for their work to
548 support our Program; Sourav Bandyopadhyay, Sergio Baranzini, Atul Butte, Gabriela
549 Fragiadakis, Julian Hong, Michael Keiser, Sara Murray, Matthew Spitzer, and Ida Sim for
550 participating as our Faculty Guest Speakers; Marilyn Ritchie for her Keynote presentation; Maya
551 Gonzalez, Nishant Jain, Anooshree Sengupta, and Carolyn Wang for participating in our
552 Undergraduate Students Panel; Stella Belonwu, Emmalyn Chen, and Caroline Warly-Solsberg

553 for participating in our Graduate Students Panel; Sarah Aerni, Tiffany Chen, Sam Gross, Katie
554 Planey, and Alex Morgan for participating in our Bio in AI Industry Panel; Arohee Bhoja, Tia
555 Jain, Christine Li, and Sachi Parikh for participating in our UCSF AI4ALL Alumni Panel; Bill
556 Lindstaedt for leading our students in a Personal Growth Session; the Altschuler and Wu Lab,
557 the Giacomini Lab, the Keiser Lab, the Kroetz Lab, the Krogan Lab, and the Witte Lab at UCSF
558 for their support of our program; Pranav Rajpurkar and Andrew Ng for granting us permission to
559 include images from CheXpert; and Edna Rodas and Amber Nolan for Administrative Support.

560 **References**

- 561 1. 8 Helpful Everyday Examples of Artificial Intelligence [Internet]. IoT For All. 2020 [cited 2021
562 Feb 2]. Available from: <https://www.iotforall.com/8-helpful-everyday-examples-of-artificial->
563 intelligence
- 564 2. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng. 2018
565 Oct;2(10):719–31.
- 566 3. Fröhlich H, Balling R, Beerenwinkel N, Kohlbacher O, Kumar S, Lengauer T, et al. From
567 hype to reality: data science enabling personalized medicine. BMC Med. 2018 Aug
568 27;16(1):150.
- 569 4. Noor P. Can we trust AI not to further embed racial bias and prejudice? BMJ. 2020 Feb
570 12;368:m363.
- 571 5. Adamson AS, Smith A. Machine Learning and Health Care Disparities in Dermatology.
572 JAMA Dermatol. 2018 Nov 1;154(11):1247.
- 573 6. West SM, Whittaker M, Crawford K. Discriminating Systems: Gender, Race, and Power in
574 AI [Internet]. AI Now Institute; 2019 Apr [cited 2021 Jan 5] p. 33. Available from: m
575 <https://ainowinstitute.org/discriminatingsystems.html>
- 576 7. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real
577 time. Lancet Infect Dis. 2020 May 1;20(5):533–4.

578 8. Akaike H. Fitting autoregressive models for prediction. *Ann Inst Stat Math.* 1969;21:243–7.

579 9. Chatfield C. The Holt-Winters Forecasting Procedure. *J R Stat Soc Ser C Appl Stat.*

580 1978;27(3):264–79.

581 10. Kröse B, Kroese B, Smagt P van der, Smagt P. An introduction to Neural Networks. 1993.

582 11. Jäger S, Cimermancic P, Gulbahce N, Johnson JR, McGovern KE, Clarke SC, et al. Global

583 landscape of HIV-human protein complexes. *Nature.* 2011 Dec 21;481(7381):365–70.

584 12. Ramage HR, Kumar GR, Verschueren E, Johnson JR, Von Dollen J, Johnson T, et al. A

585 Combined Proteomics/Genomics Approach Links Hepatitis C Virus Infection with Nonsense-

586 Mediated mRNA Decay. *Mol Cell.* 2015 Jan 22;57(2):329–40.

587 13. Eckhardt M, Zhang W, Gross AM, Von Dollen J, Johnson JR, Franks-Skiba KE, et al.

588 Multiple Routes to Oncogenesis Are Promoted by the Human Papillomavirus-Host Protein

589 Network. *Cancer Discov.* 2018 Nov;8(11):1474–89.

590 14. Batra J, Hultquist JF, Liu D, Shtanko O, Von Dollen J, Satkamp L, et al. Protein Interaction

591 Mapping Identifies RBBP6 as a Negative Regulator of Ebola Virus Replication. *Cell.* 2018

592 Dec 13;175(7):1917-1930.e13.

593 15. Shah PS, Link N, Jang GM, Sharp PP, Zhu T, Swaney DL, et al. Comparative Flavivirus-

594 Host Protein Interaction Mapping Reveals Mechanisms of Dengue and Zika Virus

595 Pathogenesis. *Cell.* 2018 Dec 13;175(7):1931-1945.e18.

596 16. Breuza L, Poux S, Estreicher A, Famiglietti ML, Magrane M, Tognoli M, et al. The

597 UniProtKB guide to the human proteome. *Database J Biol Databases Curation* [Internet].

598 2016 Feb 19 [cited 2021 Feb 16];2016. Available from:

599 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4761109/>

600 17. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely

601 available Python tools for computational molecular biology and bioinformatics.

602 Bioinformatics. 2009 Jun 1;25(11):1422–3.

603 18. Cohen JP, Morrison P, Dao L. COVID-19 Image Data Collection. ArXiv200311597 Cs Eess

604 Q-Bio [Internet]. 2020 Mar 25 [cited 2021 Feb 3]; Available from:

605 <http://arxiv.org/abs/2003.11597>

606 19. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. CheXpert: A Large Chest
607 Radiograph Dataset with Uncertainty Labels and Expert Comparison. Proc AAAI Conf Artif
608 Intell. 2019 Jul 17;33(01):590–7.

609 20. Lim J, Jeon S, Shin H-Y, Kim MJ, Seong YM, Lee WJ, et al. Case of the Index Patient Who
610 Caused Tertiary Transmission of Coronavirus Disease 2019 in Korea: the Application of
611 Lopinavir/Ritonavir for the Treatment of COVID-19 Pneumonia Monitored by Quantitative
612 RT-PCR. J Korean Med Sci [Internet]. 2020 Jan 8 [cited 2021 Feb 25];35(6). Available from:
613 <https://doi.org/10.3346/jkms.2020.35.e79>

614 21. COVID-19 pneumonia | Radiology Case | Radiopaedia.org [Internet]. [cited 2021 Feb 25].
615 Available from: <https://radiopaedia.org/cases/covid-19-pneumonia-7>

616 22. Mick E, Kamm J, Pisco AO, Ratnasiri K, Babik JM, Calfee CS, et al. Upper airway gene
617 expression differentiates COVID-19 from other acute respiratory illnesses and reveals
618 suppression of innate immune responses by SARS-CoV-2. medRxiv. 2020 May
619 22;2020.05.18.20105171.