To: Graduate Admissions Committee
From: Manolis Kellis, Ph.D.

Dear Members of the review committee,

I am happy to provide a recommendation for Hans (Wenhao) Han for your program. I was Hans's teacher for an extracurricular online educational/research program in July-August 2021, with lectures, meetings, and mentoring sessions over the span of 6 weeks.

Below, I provide a detailed evaluation of Hans in the course (section 1), a detailed description and evaluation of the project that Hans carried out in a team of students (section 2), a detailed description of the course and its structure (section 3), and more details on my own background (section 4).

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. Detailed Student Evaluation for Hans

Hans was a very good student in the class, with great programming and data mining skills. He was a driving force behind one of the most successful projects in the class and worked very well with his team. He was mainly responsible for data and model integration, gathering and processing datasets, and testing and running code on real datasets, and he made great contributions that were crucial for his team. In addition, he always asked valuable questions for both project guidance and teaching materials during lecture and TA sessions, demonstrating great communication and intellectual skills.

Out of 25 outstanding, self-selected, smart, well-prepared, and hard-working students that participated in the course, Hans **was ranked #11**, earning an **overall letter grade of A**, with an average score of 90/100. Hans's team was ranked 4 out of 8, and Hans within-team contribution was 1 (1=best, 9=worst).

Hans carried out a project titled "***Classifying Autism Spectrum Disorder Using Machine Learning through ABIDE Dataset***" as part of Team E. The team received a 5.5/5.0 overall score, with individual scores of 5.2 for Originality/Innovation, 5.4 for Challenge (Absolute), 5.3 for Challenge (Relative to their team's background), 5.4 for Problem Importance, 5.6 for Relevance, 5.4 for Slides, 5.1 for Oral Presentation, 5.4 for Written Report, 5.0 for Achievement, and 5.7 for Teamwork/Coordination. **For each of these metrics, a score of 5.0 is meeting all requirements for completing the course with an excellent project. Any score above 5 demonstrates truly exceptional work.**

You can find a **detailed description and evaluation** of Team E's project in **section 2 below**, with their abstract, links to their final report, final presentation slides, and the video of their final presentation, as well as detailed scores by the staff. Please look through these materials, as they truly allow you to judge both the overall project and Hans's contribution in great detail.

Hans was **21 years old** when taking this course, which is very young for such complex material combining algorithms, machine learning, statistics, genomics, and human health. Hans was enrolled at **Xi'an University of Architecture and Technology** (Major: **Communication Engineering**, Year: **University Year 3 (Junior)**). Hans's motivation for taking the class was: "*I want to learn AI and deep learning algorithm. I also want to gain some research experience.*" and stated career goals were: "*I want to apply for postgraduate education in the UK or Hong Kong.*".

Hans's self-reported background **at the beginning of the course** was:

- In **algorithms** 0/5, writing "*None Reported*"
- In **programming** 1/5, writing "*None Reported*"
- In **machine learning** 1/5, writing "*None Reported*"
- In **biology** 1/5, writing "*None Reported*"
- In **mathematics** 2/5, writing "*None Reported*"
- In carrying out independent **research** 0/5, writing "*No*"
- **English Language**: Listening (understanding) fast presentations: 1/5; Speaking / presenting: 1/5; Reading technical topics: 1/5; Writing technical topics: 1/5.

I provide more details on **Hans's specific contribution to the team project**, and the contributions of each student in Team E after describing the project itself in the next section.

2. Hans's Project Evaluation (as part of Team E)

Project title: Classifying Autism Spectrum Disorder Using Machine Learning through ABIDE Dataset

Team members: Yaluo, Stephanie, Hans

Abstract: *Autism spectrum disorder (ASD) is a heterogenous neurodevelopmental disorder which is notoriously difficult to diagnose, especially in children. The current psychiatric diagnostic process is based purely on the behavioral observation of symptomology (DSM-5/ICD-10) with poor understanding of the neurological mechanisms underlying ASD, and may be prone to misdiagnosis. In order to move the field toward more quantitative diagnosis, we need an advanced and scalable deep learning infrastructure that will allow us to identify reliable biomarkers of mental health disorders. This project will explore a deep learning architecture as well as some other machine learning methods for classifying patients with ASD from typical control subjects using the fMRI data. The ultimate goal of this project will be to help advance our understanding of the neurobiological underpinning of the ASD brain. Furthermore, the method used in the project can lead to accurate and early detection of ASD.*

Links to Team E final report, final slides, and final presentation video recording:

- Final report: https://www.dropbox.com/s/1pwiv4085vdbjod/TeamE_FinalReport.pdf?dl=0
- Final slides: https://www.dropbox.com/s/y5r0rj4rxcmk2sp/TeamE_PresentationSlides.pdf?dl=0
- Video Recording of final presentation: https://youtu.be/wwxw8cQ7YgI

Short Description: Team E research project topic is Autism spectrum disorder. They seek to improve the accuracy of Autism Spectrum Disorder diagnosis by using machine learning approaches to classify the fMRI data and finally increased the result accuracy of previous papers Specifically, they used a linear data augmentation method to expand the dataset, by mixing up two random samples, and use K-nearest neighbor to generate new data samples. For the architecture part of the model, they tried to use the sparse autoencoder and variational auto-encoder to learn the latent representation of the data and then designed the deep learning model and the ensembled machine learning model to finish the classification task. Finally, they achieved better result accuracy increased from 70.8% of the previous paper to 75.5%. Overall, they successfully found a machine learning architecture to help accurate and early detection of ASD.

Take-Home: The team generated the idea of using the ensembled learning method in the final classification model, which the previous paper wasn't using. Moreover, they learned a lot of machine learning approaches and models such as VAE, Sparse auto-encoder, DNN, and so on. Practically, they used different algorithms and neural network architectures to find out how they perform, and which one might work the best.

Most Impressive: Team E finished the project with good results and the division of labor was very good, which took the advantage of every single member. The data augmentation technique they try and used like mixup and dimension reduction is very promising. Moreover, the understanding and building of machine learning models based on different algorithms are very impressive.

To Improve: Team E tried a large number of methods more than ten or twenty, but each model got similar results, so it is necessary to consider whether there is a problem with predecessor data preprocessing. Sometimes the data determine the best performance (upper limit) of the model, and these cannot be improved from the model. The team may need to extract the unusable data, except for noise and outliers.

Average scores for Team E:

Team Rank	Ranked #4 out of 8 teams
1.Originality/Innovation	5.2/5.0
2.Overall challenge (absolute)	5.4/5.0
3.Relative challenge (scaled to team)	5.3/5.0
4.Problem importance	5.4/5.0
5.Relevance	5.6/5.0
6.Slides	5.4/5.0
7.OralPresentation	5.1/5.0
8.Written report	5.4/5.0
9.Achievement	5.0/5.0
10.Teamwork	5.7/5.0
Overall Team Score	5.5/5.0

Hans's Detailed Contributions to different aspects of Team E's project:

- **Overall**: In our project, I mainly responsible for gathering the datasets, applying different methods for data augmentation, integrating and testing the code of our model. I also participated in generating the results for visualization and modifying our methods. Besides , I took part in nearly all paperwork including proposal, end-to-end pipline, final report and slides.

- **1. Idea/Innovation**: I applied spearman correlation for approximating the functional connectivity of different brain regions. Also, I augmented our insufficient datasets with linear interpolation method (mixup) and achieved better classification accuracy. I also tried some nonlinear augmentation methods including VAE and t-SNE but didnt achieve better result.
- **2. Methods/Programming Contributions**: I modified our code in terms of the correlation method. And I wrote and debug the code of mixup(our linear interpolation method), tested its performance in our datasets. Finally, I combined the code of Sparse AutoEncoder and Variation AutoEncoder with DNNs, run and debugged the code to generated results for visualization.
- **3. Analysis/Results/Interpretation Contributions**: After combining all methods together, I tested our model on ABIDE datasets (consists of data form 17 different sites) using 10-fold cross-validation and analyzed the results. Then, I tested our model on data of each site using 5-fold cross-validation and evaluated its performance on ASD classification.
- **4. Report/Presentation Contributions**: As for our final report, I wrote the part of feature extraction and data augmentation. As for our slides, I made the part of fMRI technique, ABIDE datasets, feature extraction, data augmentation method and its results. As for our video presentation, I explained our datasets, data preprocessing method, the workflow of feature extraction and the data augmentation method.
- **5. Other Contributions**: I interpreted the datasets and the methods in the paper to my teammates. I also helped my teammate debug the code in their parts.

Contributions for each **other** student in Hans's team:

Yaluo's Detailed Contributions to different aspects of Team E's project:

- **Overall**: Even though there is no leader's role within the team, I take the most lead in making executive decisions and helping to arrange group meetings to proceed with the research. For the implementation of the model, I tried to assist by adding the Variational Autoencoder method into the Deep Neural Network. In terms of the report and slides, I initiated the general format. I make all the data visulization and wrote in detail on the Sparse Autoencoders, Variational Autoencoders, Result & Evaluation, Division of Labors, Comment on experience, Comment on peer review, Reference.
- **1. Idea/Innovation**: The project idea is based on the first paper I chose to review for this class. Though I wish there could be more time for the team to decide on the topic, we still settle down on the Autism Spectrum Disorder classification. The silver lining is that I am very passionate about the subject and have done past projects around Autism disease in design and art forms. With additional research on the topic, we found more papers like a family structure"father, son, grandson," since the papers are closely connected and trying to improve the final outcome based on each other. I also traced down the relevant codes and ABIDE dataset that is the foundation of the project later.
- **2. Methods/Programming Contributions**: Little did I have any experience in Deep Neural Network coding construction or data augmentation. Therefore, when I learned about the autoencoders in class, I immediately thought it could be one innovation for the project. Then, I help with the coding of Sparse Autoencoder and Variational Autoencoder, leaving the other code to the expertise in the group. Due to the limitation of GPUs that Colab is available to give for each user, I later help with the running of code to generate more valuable results.
- **3. Analysis/Results/Interpretation Contributions**: The model architecture that the team used is heavily dependent on GPUs. Each of us has helped to generate the results for the evaluation. I did the part to compare and contrast the outcome between Mixup+Variational Autoencoders + Deep Neural Network with "Mixup+ Sparse Autoencoders + Deep Neural Network." Moreover, I create labels to document the differences between Team E and previous studies' methods.
- **4. Report/Presentation Contributions**: As for the presentation aspects, I always try to make a little structure for the teammates to follow later. I initiate the visual format and outline for the report and the slides. Then, each of us has different responsible sections, so we devote ourselves to writing our parts. My part is making all the data visualizations not limited to tables, graphs, and flowcharts. I have also elaborated on the report and slide description of the sections: Sparse Autoencoders, Variational Autoencoders, Result & Evaluation, Division of Labors, Comment on experience, Comment on peer review, Reference. After recording the video, I try to edit through Adobe Premiere hoping to submit the best oral presentation as a team.
- **5. Other Contributions**: I guess I am taking the leadership quite often in the process of this summer research. I also tried to bring out the past research experiences from before and see what the team could adapt. I assist the team in making better oral presentations and encouraging them to speak more in class. It is my honor to be on the same team as Stephanie and Hans. I am delighted to have worked with them and accomplished this team research project.

Stephanie's Detailed Contributions to different aspects of Team E's project:

- **Overall**: In terms of model, I'm responsible for working out the algorithm implementation, especially the model training and the possible ensemble learning methods and deep neural network modification. In terms of the writing of the proposals, I focused on the Introduction, Related Work, Significance, Innovation, ensemble learning and deep neural networks in the Approach. And in the process of writing the final report, I focused on similar sections that I have written above.
- **1. Idea/Innovation**: I mainly figured out how to download the datasets.
- **2. Methods/Programming Contributions**: I wrote the code of new methods like Gradient Boosting classifier and K-Nearest Neighbors classifier. And I combined the original 3 models with the two classifiers using ensemble learning methods. I also replaced the classifier used in the 'ASD-DiagNet' with a deep neural network.

- **3. Analysis/Results/Interpretation Contributions**: I compared the predicted lables of the DiagNet,SVM,RandomForest,Gradient boosting classifier and K-nearest Neighbors classifier in an excel sheet. This procedure can help us figure out why the ensemble learning methods didn't work in our project.
- **4. Report/Presentation Contributions**: In terms of the writing of the proposals, I focused on the Introduction, Related Work, Significance, Innovation, ensemble learning and deep neural networks in the Approach. And in the process of writing the final report, I focused on similar sections that I have written above. In the making of the slides and video presentation, I was responsible for the parts of the related work, project aims and motivation, model architecture, and two autoencoders with deep neural netwroks.
- **5. Other Contributions**: I also applied my knowledge to help other team members with their parts like helping with the understanding of the original paper, the change of the correlation calculating methods.

3. Course description

The course was offered remotely (over Zoom) over 6 weeks, meeting on Saturday and Sunday for 4 weeks, followed by meeting daily for 2 weeks, from July 3 - August 6, 2021. It was taught by Prof. Manolis Kellis, with TAs Zijian Wang and Simiao Zhao.

This course introduced foundational and state-of-the-art machine learning techniques in the context of understanding the human genome and human disease mechanism. It covered both the computational foundations and the research frontiers of the field of computational biology.

Machine learning, statistical, and algorithmic techniques included: Bayesian inference, deep learning, hidden Markov models (HMMs), random forests, convolutional neural networks, auto-encoders, recurrent neural networks, clustering and classification, k-means, hierarchical clustering, model complexity selection, network structure, PCA, SVD, network diffusion kernels, quantitative trait mapping, mediation analysis, causality inference, string matching, sequence alignment, tree data structures, rapid database search, hashing, data integration, pattern finding, expectation maximization, Gibbs sampling.

Biological applications included: Genetic association mapping, common/rare variants, GWAS, PheWAS, multi-trait mapping, EHR mining, cancer genomics, CRISPR, biological sequence analysis, gene finding, comparative genomics, RNA structure folding, sequence alignment, gene expression analysis, motifs, epigenomics, single-cell genomics, evolutionary analysis, gene/species trees, phylogenomics, coalescent, personal genomics, population genomics, human ancestry, recent selection, disease mapping, genetic association analysis, population genetics, regulatory genomics, dissecting disease mechanism.

The course was organized as follows:

Date	Week	Overall topic	Day	Individual Topics
	Before	Why we do all this	0	Day 0 = Introduction to research in genomics - Research Lectures by Prof. Manolis Kellis
Sat, July 3	Week 1	Intro + Machine Learning	1	Day1.1=Course Intro, Biology, Algorithms, Machine Learning, Project Overview
				Day1.2=Machine Learning Foundations, Supervised Learning, Bayesian Inference, Clustering, Classification, K-means
Sun, July 4		Deep Learning: CNNs, RNNs	2	Day2.1=Deep Learning, Neural Nets, Convolutional NNs, Representation Learning, Autoencoders
				Day2.2=Recurrent NNs, Graph NNs, Generative Models, Representations, Interpretability
				Project Milestone 1: Self-Introduction, Video #1 Recording on self-presentation, 1-page Sheet, Types of Projects
Sat, July 10	Week 2	Graphs,Netwks, Embeddings	3	Day3.1=Linear Algebra, Networks, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels
				Day3.2=Embeddings, t-SNE, NNMF, Dimensionality Reduction, SVMs, LDA, KNNs, Random Forests, Regression
Sun, July 11		DL2: GNNs, GANs, VAEs	4	Day4.1=Deep Learning 3: Graph Neural Networks (GNNs), Variational AutoEncoders (VAEs), Generative Adversarial Networks (GANs)
				Day4.2=Deep Learning 4: Interpretability, Representation Learning, Drug Design, Protein Folding
				Project Milestone 2: Review previous papers, previous projects, available datasets, methods, **Report + Video #2 on previous papers**
Sat, July 17	Week 3	Algorithms, DynProg, HMMs	5	Day5.1=Algorithms, Dynamic Programming, String Search, Scheduling, Hashing, BWT transforms, Database Search, Suffix Search
				Day5.1=Probabilistic Models, HMMs1: Evaluation, Parsing, Posterior Decoding, Learning, HMM architectures, CRFs
Sun, July 18		Genomics, Epi/Tx, Single-cell	6	Day6.1=Genomics, DNA, Gene Regulation, Epigenomics, Regulatory Genomics
				Day6.2=Single-cell genomics, tSNE, Cell Type Annotation, scRNA/scATAC, linking
				Project Milestone 3: Project Pre-proposal (Individual): Brainstorm Project Ideas, Write-up + **Video #3** Recording on Project Preproposal
Sat, July 24	Week 4	Variation,Disease,Neur/Canc	7	Day7.1=Human Genetics, PopGen, LD, Disease Association Mapping, GWAS, Complex Traits
				Day7.2=Electronic Health Records (EHRs), Clinical data text mining, Biological Image Analysis, Cancer Detection, Neuroscience
				Project Milestone 4: Team Formation and General Project Topics Selection, write-up + **First Team Joint Video Recordings #4**
Mon, Jul 26	Week 5	Proposal finalization	8	Milestone 5: Gather more papers for selected topics, datasets, code, Brainstorm integrative projects, key questions, missing components
Tue, Jul 27			9	Milestone 6: **Formal Project Proposal** (team): Aims, Milestones, Timeline, Alternative Strategies. **Video Recording: Project Proposal #5**
Wed, Jul 28		Data Exploration	10	Milestone 7: **End-to-End Pipeline video #6**: Make sure all datasets are available, code runs, demonstrate parsing, data visualization.
Thu, Jul 29			11	Milestone 8: Exploratory Analyses, Implement Learning Methodologies, Ask biological/medical Questions, Formulate Hypotheses
Fri, Jul 30		Midcourse report	12	Milestone 9: Planning for the mid-course report: What should be in the final report, walking back from the final report to what is missing
Mon, Aug 02	Week 6		13	Milestone 10: Mid-course report: The known unknowns, and the unknown unknowns of what remains to be done. **Video #7**
Tue, Aug 03		Visualization, Flow	14	Milestone 11: Completing Analyses, Finalizing Figures and Tables, Writing Introduction and Methods. Abstract and Conclusions.
Wed, Aug 04			15	Milestone 12: Making final figures, writing figure legends, visual legends, take-Home messages, Organizing your slide deck #1
Thu, Aug 05		Writing and Oral presentations	16	Milestone 13: Polishing reports, Polishing Figures, Slide organization, Presentation Preparation. **Draft Presentation Video #8**
Fri, Aug 06			17	Milestone 14: Final reports due, Pre-Final presentation **Video #9**. Final feedback on report & presentation. Final versions due (**v#10**)
				Grading of final submissions, letters released, letters will follow over the next few days

In addition to the technical material in the course, a course-long final term project enabled students to design, plan, carry-out, and present their own independent research projects, and to become active practitioners in the field of computational biology. A series of exercises, guidance, advice lectures, and mentoring sessions throughout the class enabled students to: (1) introduce themselves and their interests and form teams; (2) design a project, identify relevant datasets and algorithms, form teams, and plan out their research; (3) write an NIH-style research proposal outlining their goals and milestones; (4) review peer proposals, provide constructive feedback, and incorporate peer feedback into their own projects; (5) design and modify software for data analysis and integration, interpret their results, and draw biological constructions; (6) present their research results orally in a conference setting; and (7) present their results in written form in a journal-style scientific paper.

4. About the instructor:

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 240 journal publications, which have been cited more than 120,000 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: compbio.mit.edu.