

Thomas Cheng Li

121 Jericho Rd
Weston, MA 02493

25lit@weston.org
TomLi.org

EDUCATION

Weston High School

GPA: 3.96/4.0 **ACT:** 35

AP Calculus BC, AP Physics C: Mechanics and E&M, AP Chemistry, AP Computer Science, AP World History, AP U.S. History, AP Psychology, AP Biology, AP European History, AP Statistics, AP Spanish, AP English

PROJECTS

Functional Annotation of Variants Online Resource (FAVOR)

Mar. 2023 - Present

Prof. Xihong Lin, Harvard University

- Aggregated large genomic data (9 billion single nucleotide variants, 80 million observed genetic insertions/deletions, and 160 columns of variant annotations) into FAVOR (<https://favor.genohub.org>), an online database widely used by medical researchers to connect disease traits with human genetic information.
- Conceptualized and developed a natural language interpretation model (FAVOR-GPT) to easily query and extract relevant information from the database and generate insights, reducing the barrier of entry and making the database more accessible to researchers worldwide.
- Presented research at The Impact of Genomic Variation on Function (IGVF) Consortium and published a **first-author paper** in peer-reviewed *Bioinformatics Advances*, 2024, [vbae143](#).

Sign Language Recognition

Jan. 2021 - Aug. 2023

https://github.com/hsteven-archive/sign_lang

- Worked with another high-schooler to build a computer vision program using Python to recognize and translate American Sign Language (ASL) into English in real-time.
- The model can be used to automatically caption and translate videos for accessibility or to teach ASL more effectively, with the goal of improving communication for the deaf and hard of hearing.
- Wrote a **co-first-author paper** published in the *Research Archive of Rising Scholars* (2023).

PUBLICATIONS

FAVOR-GPT: A Natural Language Interface to Genomic Annotations

Thomas Cheng Li, Hufeng Zhou, Vineet Verma, Xiangru Tang, Yanjun Shao, *Bioinformatics Advances* (2024)
Eric Van Buren, Zhiping Weng, Shamil R Sunyaev, Mark Gerstein, Xihong Lin doi.org/10.1093/bioadv/vbae143

Sign Language Recognition from Video using Geometrical and Transfer Learning Techniques

David Chen and Thomas Cheng Li (Co-first authors)

Research Archive of Rising Scholars (2023)

doi.org/10.58445/rars.486

EXTRACURRICULARS

Princeton University Laboratory Learning Program (LLP)

Summer 2024

Utilizing LLMs to find information on Wastewater Treatment Facilities, with Prof. Jason Ren

- 6% acceptance rate summer internship experience to work under accomplished Princeton professor.
- Built program that is able to automatically gather publically available information for wastewater treatment plants and other environmentally related facilities.
- Expanded the capabilities of state of the art LLMs for answering environmental engineering questions using vector databases.

Stony Brook University Simons Summer Research Program (SSRP)

Jul. 1 - Aug. 9, 2024

Characterizing Machine Learning Methods for Medical Time Series Diagnosis, with Dr. Alisa Yurovsky

- 5% acceptance rate summer research program to do bioinformatics research at Stony Brook University.
- Trained time series transformer classification models in order to detect Acute Kidney Failure using Electronic Health Records from the Trinetx database.
- Analyzed and tested different data imputation methods for patient data to increase model performances.

- Trained Generative Adversarial Networks to generate synthetic EHR data that balances underrepresented patient groups or conditions, reducing bias and improving the model's fairness and generalization across diverse populations.

Physics Club, Weston High School

Sep. 2022 - Present

Co-President

- Founded high school Physics Club by gathering 25 interested student signatures, finding a faculty advisor, and presenting idea to administrators for approval.
- Organized club meetings every other week to practice and discuss challenging physics problems, tutor Freshman students for Honors Physics.

Cybersecurity Team, Weston High School

Oct. 2021 - Present

Captain

- Competed in Cyberpatriot's National Youth Cyber Defense Competition, held by the Air & Space Forces Association where thousands of students from around the US compete in securing compromised Windows/Linux systems; Won the Platinum Division 1st Place State Award and the Platinum Division Semifinalist Award.
- Competed in the Cyberstart program where 42,402 students competed to solve challenging security-related problems; Won the National Semifinalist award.
- Elected team captain and organized weekly meetings for ~10 student members to practice cybersecurity competition problems, published 132-page book *Digital Literacy 101* on amazon.com/dp/B0DBSWX8ZL to which I contributed five chapters.
- Interned part-time for cybersecurity firm BG Networks, providing security solutions against ransomware.

Students for Environmental Action (SEA) group

Oct. 2022 - Present

Co-President

- Volunteered 2 hours every week to support local sustainability by planting trees, building forest trails, and encouraging recycling (bought and set up 5 of the first recycling trash cans in school history).
- Lobbied and testified in front of state congress over climate bills (e.g. Air Quality EJ bill), engaged in local town sustainability decisions as a student representative on the sustainability committee, and joined the Massachusetts Youth Climate Coalition (MYCC), aiming to tackle air quality control, voting transparency, and climate-friendly building renovations. A member of [Town of Weston's Sustainability Committee](#).
- Elected Vice-President and organized weekly meetings, volunteering opportunities, and other initiatives amongst the ~7 consistent members. Built and maintained its website seaweston.org

Visitor Volunteer at NewBridge on the Charles Senior Home

Apr. 2024 - Present

AWARDS

United States of America Computing Olympiad (USACO) Gold (testing for Platinum) 2024

Certificate of Distinction (placed [133rd in the world out of 3,077 participants](#)) in Sir Isaac Newton Exam of Physics, organized by the Department of Physics & Astronomy, University of Waterloo 2024

Cyberpatriot's National Youth Cyber Defense Competition 2023

Platinum Division 1st Place State Award, Platinum Division Semifinalist Award, Gold Division 1st Place Award.

- The nation's largest cyber defense contest (over 3,000 schools participating with 5 members per school on average), held by the Air & Space Forces Association.
- Received a [letter of congratulation from Massachusetts State Representative Alice Peisch](#).

3rd Place Award at the [Massachusetts Science & Engineering Fair \(MSEF\)](#) 2024

Analysis and Machine Learning Modeling of Spatial Data to Identify Asthma Hotspots in Massachusetts

Best Research Project (2nd place) Award for \$1,500 cash prize (team) 2023

AAASE Princeton University Summer Academy

National Merit Scholarship Semifinalist (received the maximum score of 1520 on the PSAT) 2023

Databases and ontologies

FAVOR-GPT: a generative natural language interface to whole genome variant functional annotations

Thomas Cheng Li^{1,2,†}, Hufeng Zhou^{1,†}, Vineet Verma¹, Xiangru Tang³, Yanjun Shao³, Eric Van Buren¹, Zhiping Weng⁴, Mark Gerstein^{5,6}, Benjamin Neale^{7,8}, Shamil R. Sunyaev^{8,9}, Xihong Lin^{1,8,10,*}

¹Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, 02115, United States

²Weston High School, Weston, MA 02493, United States

³Department of Computer Science, Yale University, New Haven, CT, 06520, United States

⁴Program in Bioinformatics and Integrative Biology, University of Massachusetts Chan Medical School, Worcester, MA, 01605, United States

⁵Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, 06520, United States

⁶Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, 06520, United States

⁷Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, 02114, United States

⁸Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, 02142, United States

⁹Department of Biomedical Informatics, Harvard Medical School, Boston, MA, 02115, United States

¹⁰Department of Statistics, Harvard University, Cambridge, MA, 02138, United States

*Corresponding author. Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, 02115, United States.

E-mail: xlin@hsph.harvard.edu

[†]Equal contribution.

Associate Editor: Alex Bateman

Abstract

Motivation: Functional Annotation of genomic Variants Online Resources (FAVOR) offers multi-faceted, whole genome variant functional annotations, which is essential for Whole Genome and Exome Sequencing (WGS/WES) analysis and the functional prioritization of disease-associated variants. A versatile chatbot designed to facilitate informative interpretation and interactive, user-centric summary of the whole genome variant functional annotation data in the FAVOR database is needed.

Results: We have developed FAVOR-GPT, a generative natural language interface powered by integrating large language models (LLMs) and FAVOR. It is developed based on the Retrieval Augmented Generation (RAG) approach, and complements the original FAVOR portal, enhancing usability for users, especially those without specialized expertise. FAVOR-GPT simplifies raw annotations by providing interpretable explanations and result summaries in response to the user's prompt. It shows high accuracy when cross-referencing with the FAVOR database, underscoring the robustness of the retrieval framework.

Availability and implementation: Researchers can access FAVOR-GPT at FAVOR's main website (<https://favor.genohub.org>).

1 Introduction

Multi-faceted variant functional annotation plays a pivotal role in the analysis and interpretation of the findings of array-based Genome-Wide Association Studies (GWAS) and WGS studies (Watanabe *et al.* 2017, Li *et al.* 2020, Quick *et al.* 2020). Examples of large scale WGS studies include the Trans-Omics Precision Medicine (TOPMed) Program, UK Biobank, and *All of Us* (Sudlow *et al.* 2015, Karczewski *et al.* 2020, Taliun *et al.* 2021). Variant function annotations can be used for functional fine mapping (Kichaev *et al.* 2014, Schaid *et al.* 2018), partitioned heritability (Finucane *et al.* 2015), polygenic risk scores (PRSs; Marquez-Luna *et al.* 2021), and rare variant association analysis of WGS studies (Li *et al.* 2022).

The Functional Annotation of Variants Online Resources (FAVOR) database and portal (Zhou *et al.* 2023) provides an open access comprehensive online platform for functional

annotations of genetic variants, genomic regions and genes across the whole genome. FAVOR efficiently summarizes and visualizes multi-faceted functional annotation data of all possible (approximately nine billion) single nucleotide variants (SNVs), and insertion and deletion variants (Indels) observed in large-scale genome sequencing studies, such as TOPMed, covering the entire human genome. It enables quick and convenient querying at variant, gene, and region levels. FAVOR integrates variant functional information from diverse sources to elucidate the functional attributes of variants, and assists the prioritization of potential causal variants influencing human phenotypes. However, effectively utilizing FAVOR necessitates a certain level of prior specialized knowledge and background. Users are required to possess a fundamental understanding of different annotation metrics and the specific genes or variants they wish to query,

Received: August 27, 2024; Revised: August 30, 2024; Editorial Decision: September 10, 2024; Accepted: September 27, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

in addition to adhering to the correct input formats. Second, there are various terms and scores that users may need to refer to the FAVOR documentation to understand (Zhou et al. 2023). Third, the queried results on the FAVOR portal are static with raw annotation results, precluding interactive calculation of summary statistics of interest.

There is a significant need to develop a user-friendly tool to respond to natural language queries, and provide results in an interactive format that are easy to understand without prior knowledge. This will help bridge the gap in accessibility and usability of variant functional annotations in genetics and genomic research. There are increasing interests in leveraging Large Language Models (LLMs; Touvron et al. 2023), such as ChatGPT, GPT-4 (OpenAI 2023) and LLaMA (Touvron et al. 2023) in biomedical applications. This transformative technology offers attractive artificial intelligence capabilities. For example, GPT-4 have shown proficiency and intelligence in human interactions, achieved through instruction tuning and feedback-based training. These potentials have ignited significant interest and excitement within the scientific research community toward LLMs (Sallam 2023). In the open-source world, LLaMA has become increasingly popular (Touvron et al. 2023). LLaMA3.1's performance is on par with GPT-4. This advancement shows great potential for researchers seeking to enhance customization. Recently, VarChat (Paoli et al. 2024) was introduced to integrate chatbot-based variant search with the publications from PubMed to generate summaries. It is, however, limited to the small subset of variants documented in the published literature. It lacks the ability to query for multi-faceted functional annotations of any variant (SNV) across the human genome, and fails to provide functional information for a large number of variants in WGS studies.

In this paper, we introduce FAVOR-GPT, an interactive tool that leverages knowledge-guided LLMs to enhance the user experience interacting with the FAVOR database. Compared to the competitive products, we selected the ChatGPT API from OpenAI for following reasons. First, it offers high-quality and contextually relevant responses, while boasting rapid response times, ensuring users receive prompt replies to their queries. Second, ChatGPT provides extensive tools available in the JavaScript ecosystem, and its support for function calling makes it an ideal candidate for adopting the Retrieval-Augmented Generation (RAG) approach. It allows to integrate external knowledge sources and our in-house FAVOR APIs seamlessly into the language model's generation framework, enhancing the accuracy and relevance of the responses. Third, opting for ChatGPT APIs eliminates the need to run a local language model, and reduces the amount of additional responsibilities and complexities, such as hardware requirements, model fine-tuning, and maintenance. ChatGPT offers a more straightforward setup process, enabling us to focus on building our applications rather than managing the underlying infrastructure.

FAVOR-GPT exhibits the ability to understand user inputs in natural language and improve user experience in navigating the FAVOR database and portal. Its inherent flexibility allows it to accommodate a wide range of input formats, ensuring that queries are properly understood. In addition to retrieve query results from the FAVOR database, FAVOR-GPT has several attractive features. When presenting raw annotation results and values, it enriches these findings with relevant background introduction and leverages the natural language explanations generated by LLM (ChatGPT). It also generates

summary statistics calculated using the FAVOR database in response to prompts. This integrative approach significantly improves the understanding of functional annotation results, making the utilization of FAVOR easier for researchers. FAVOR-GPT introduces a practical approach to integrating LLMs specifically tailored for variants functional annotation, without the substantial resource requirements of pretraining or fine-tuning large models.

2 Methods

FAVOR-GPT was developed with flexibility, resource efficiency, and adaptability in mind. It was made to combine the documentation information with the annotations from the FAVOR database.

FAVOR-GPT is implemented based on the Retrieval Augmented Generation (RAG; Guo et al. 2023) approach, an AI framework that enhances responses based on an external textual knowledge source. In this case, based on the text query of the user, FAVOR-GPT allows ChatGPT to retrieve textual data in real time from the FAVOR database automatically via the FAVOR API, and thus grounding the LLM on the information from the FAVOR database and documentation and related sources for generating reliable and detailed responses, see Fig. 1. For all gene-related information, FAVOR-GPT utilizes a vector database based on Weaviate to fetch information relevant to the query. Gene information is separated into categories, such as pathway, function, identification, and embedded separately using the “text-embedding-3-small” vectorizer model from OpenAI.

To enhance user comprehension of the annotation results, FAVOR-GPT employs an in-depth analysis of relevant documentation, aligning it with the values obtained from the queries. FAVOR-GPT then employs the ChatGPT APIs to generate natural language explanations of the annotation results from the FAVOR API queries, presenting the information in a format that is easy to understand. Further, FAVOR-GPT can conduct data analysis in response to various queries, such as calculating the number of pathogenic variants in BRCA1. FAVOR-GPT also allows for the user to easily cross-verify the data given in FAVOR-GPT with the database itself. All FAVOR-GPT and FAVOR API documentation can be found at <https://docs.genohub.org/>.

The workflow of FAVOR-GPT is illustrated in Fig. 1. By harnessing the natural language generation capabilities of ChatGPT, FAVOR-GPT ensures that users receive not only raw annotation data but also contextual and coherent explanations of multi-faceted functional annotations of variants, genes and genomic regions. The current version of FAVOR-GPT is implemented using the TypeScript programming language (Bierman et al. 2014) with the Vercel.AI SDK (Grammel et al. 2023), with a deliberate effort to smoothly integrate it into the existing FAVOR user interface which build on the React/Next.js framework. FAVOR-GPT's presence in the user interface is marked by a clickable floating button placed on the FAVOR website. The source code for the site can be found at <https://github.com/zhoushufeng/FAVOR-GPT>.

We conducted benchmark testing on FAVOR-GPT by randomly selecting contexts from the vector database and using GPT-4 to create 100 questions based on specific categories like gene function or location. These questions are then answered by both FAVOR-GPT and GPT-4. The answers are assessed based on two metrics: *relevance* and *accuracy*.

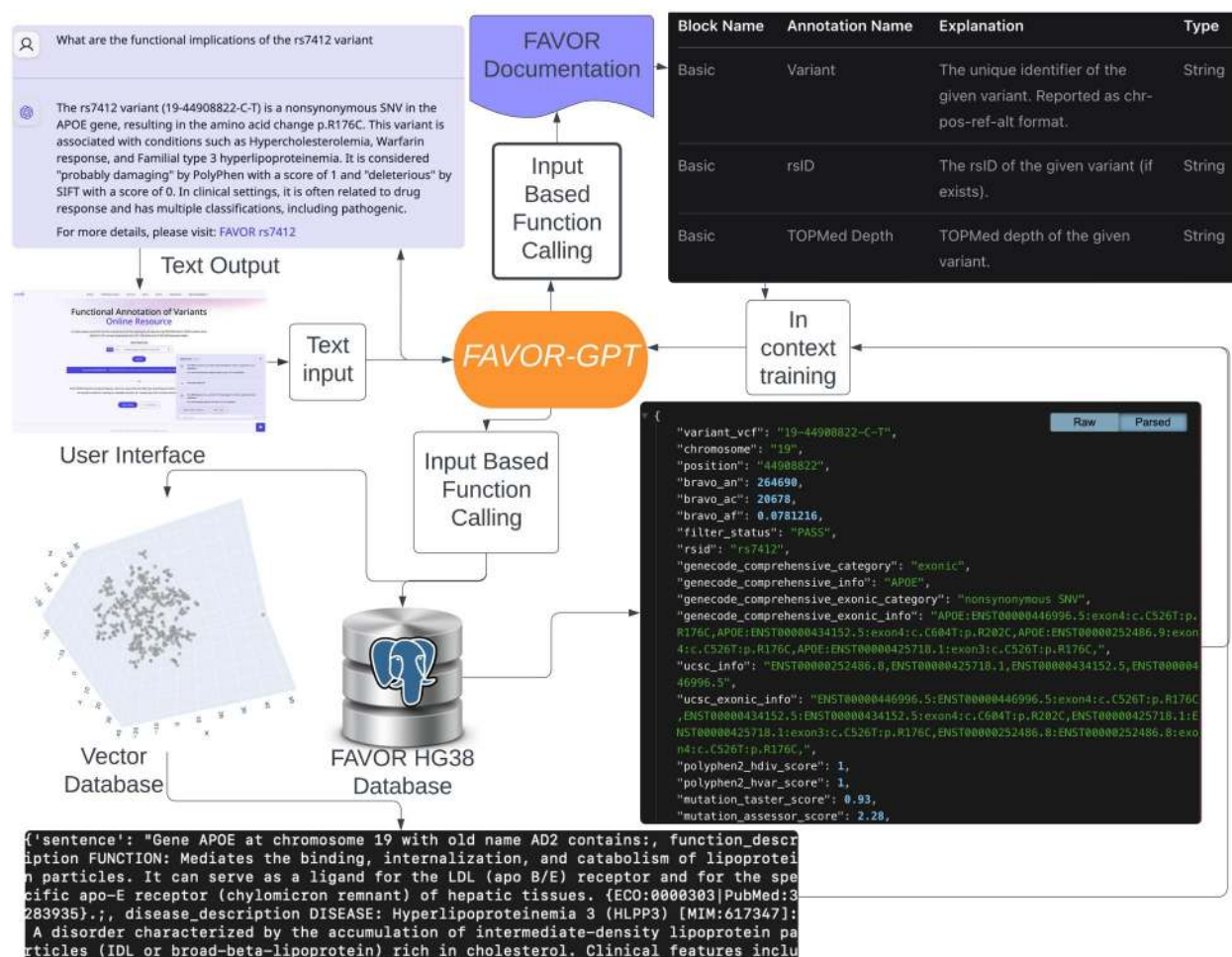


Figure 1. Graphical representation of the FAVOR-GPT workflow. The FAVOR-GPT workflow demonstrates how it converts natural language into structured query syntax and then interprets the query results into clear and fluid natural language.

Relevance measures how well the model's response addresses the question, with scores of 1 (answer directly pertains to the question), 0.5 (answer tangentially pertains to the question), and 0 (answer does not address the question at all). Accuracy measures how factually correct the answer is, with scores of 1 (completely correct), 0.5 (has mistaken but is largely correct), and 0 (factually incorrect). These scores are determined by a GPT-4 model with access to all the necessary context. To compare the model result, we had a plain GPT-4 model answer the same questions and be evaluated similarly.

3 Results

FAVOR-GPT can make any query to satisfy the text inputs. These queries include gene-level functional annotation queries, gene-based variant queries, and variant-specific functional annotation queries. [Supplementary Figure S1](#) shows examples of queries and responses. Users can ask free-form questions like "What is the function of the gene APOE?" and "What is the function of rs942096275?" FAVOR-GPT will provide comprehensive easy-to-understand answers.

FAVOR-GPT is equipped to address computational queries such as analyzing and summarizing data, for instance, gene-level and region-level variant calculations using the FAVOR database. Examples of such computational queries include "What is the range for TP53 gene?," "How many variants in APOE?," "How many pathogenic variants in BRCA1?," "How

many loss of function variants in APOE?," "How many variants in APOE with aPC Epigenetics Repressed > 20?." These responses are shown in [Supplementary Figure S2](#) and [Table S1](#). These gene-level variant calculations are performed using the TOPMed Bravo variant list, which contains observed variants in TOPMed-BRAVO and is part of the FAVOR database. This is achieved through the FAVOR API, which is designed to handle such specific queries. The FAVOR web interface offers limited gene and region level summary statistics. In contrast, FAVOR-GPT is much more flexible, enabling users to calculate a wide range of customized summary statistics based on their specific queries.

The evaluation of FAVOR-GPT shows good performance in providing variant functional annotation information. FAVOR-GPT had a relevance score of 0.865 and an accuracy score of 0.85, whereas the regular GPT-4 model had a relevance score of 0.5 and an accuracy score of 0.595 (All the examples are placed in [Supplementary Table S3](#)). In many cases, the GPT-4 model resorted to saying that it did not know the answer to the question, such as "How many pathogenic variants does BRCA1 have?" which raised the accuracy score to be decently high as "I don't know." Although the scores show that FAVOR-GPT still has room for improvement, they also show that the current RAG system by integrating the high quality whole genome variant annotation database FAVOR significantly improves gene-related and variant-related queries and calculations.

4 Discussions

We have developed FAVOR-GPT, an interactive interface that leverages Language Model APIs with the multi-faceted variant functional annotation database. It furnishes encompassing annotation results within the FAVOR ecosystem, ensuring that users have access to comprehensive knowledge-guided information and explanations. FAVOR-GPT exhibits relevance and accuracy in interpreting users' natural language inputs, translating them into structured database queries, and explaining annotation results in natural language and hyperlinks of the sources. Serving as the one of the core interfaces for accessing functional annotation within FAVOR, it is also capable of performing various summary statistics calculations using the data in FAVOR.

The utilization of FAVOR-GPT enables a wider community of researchers to more easily conduct genetics and bioinformatics research. Our efforts to harness the power of Language Model APIs to enhance bioinformatics database usage will be helpful for similar developments in the field. The advent of DNN-driven LLMs represents a valuable force for a new type of interface that improves database user experience. FAVOR-GPT sets an example for navigating large and complex databases of a similar nature. By providing a model for developing and implementing intuitive, natural language-driven interfaces, FAVOR-GPT showcases an effective implementation approach for other specialized knowledge bases to broaden their reach and enhance user experience.

Supplementary data

Supplementary data are available at *Bioinformatics Advances* online.

Conflict of interest

B.M.N. is on the Scientific Advisory Board of Deep Genomics, a consultant for Camp4 Therapeutics, Takeda Pharmaceutical and Biogen. S.R.S. is consultant to NGM Biopharmaceuticals and Inari agriculture. He is also on Scientific Advisory Board of Veritas Genetics. X.L. is a consultant of AbbVie Pharmaceuticals and Verily Life Sciences. Z.W. co-founded and serves as a scientific advisor for Rgenta Inc.

Funding

This work was supported by the National Institutes of Health [grant numbers R35-CA197449, R01-HL163560, U01HG012064, U19-CA203654, and P30 ES000002 to T.C. L., H.Z., V.V., and X.L.]

Data availability

The data and software of FAVOR-GPT underlying this article are available in FAVOR database, at <https://favor.genohub.org/> and source code of FAVOR-GPT can be accessed at <https://github.com/zhohuhufeng/FAVOR-GPT>.

References

- Bierman G, Abadi M, Torgersen M. Understanding typescript. In: *Proceedings of the 28th European Conference on ECOOP 2014 — Objective-Oriented Programming*, Uppsala, Sweden, July 28–August 1, 2014. Vol. 8586. Heidelberg: Springer-Verlag Berlin, 2014;257–81.
- Finucane HK, Bulik-Sullivan B, Gusev A *et al.*; RACI Consortium. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* 2015;**47**:1228–35.
- Grammel L, Leiter M, Palmer J *et al.* Vercel AI SDK. The Vercel AI SDK is a library for building AI-powered streaming text and chat UIs. 2023. <https://github.com/vercel/ai>
- Guo Y, Qiu W, Leroy G *et al.* Retrieval augmentation of large language models for lay language generation. *J Biomed Inform* 2023; **149**:104580.
- Karczewski KJ, Francioli LC, Tiao G *et al.*; Genome Aggregation Database Consortium. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;**581**:434–43.
- Kichaev G, Yang W-Y, Lindstrom S *et al.* Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet* 2014;**10**:e1004722.
- Li X, Li Z, Zhou H *et al.*; TOPMed Lipids Working Group. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat Genet* 2020;**52**:969–83.
- Li Z, Li X, Zhou H *et al.*; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium. A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nat Methods* 2022;**19**:1599–611.
- Márquez-Luna C, Gazal S, Loh P-R *et al.*; 23andMe Research Team. Incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. *Nat Commun* 2021; **12**:6052.
- OpenAI. ChatGPT. 2023. <https://openai.com/> (6 June 2024, date last accessed).
- Quick C, Wen X, Abecasis G *et al.* Integrating comprehensive functional annotations to boost power and accuracy in gene-based association analysis. *PLoS Genet* 2020;**16**:e1009060.
- Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)* 2023;**11**:887.
- Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* 2018;**19**:491–504.
- Sudlow C, Gallacher J, Allen N *et al.* UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;**12**:e1001779.
- Taliun D, Harris DN, Kessler MD *et al.*; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 2021; **590**:290–9.
- Touvron H, Martin L, Stone K *et al.* Llama 2: Open foundation and fine-tuned chat models. arXiv, arXiv:2307.09288, 2023, preprint: not peer reviewed.
- Watanabe K, Taskesen E, van Bochoven A *et al.* Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* 2017;**8**:1826.
- Zhou H, Arapoglou T, Li X *et al.*; NHGRI Genome Sequencing Program Variant Functional Annotation Working Group. FAVOR: functional annotation of variants online resource and annotator for variation across the human genome. *Nucleic Acids Res* 2023; **51**:D1300–11.

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Bioinformatics Advances, 2024, 00, 1–4

<https://doi.org/10.1093/bioadv/vbae143>

Application Note

FAVOR-GPT: A Generative Natural Language Interface to Variant Functional Annotations

Supplementary Materials

Thomas Cheng Li^{1,2,*}, Hufeng Zhou^{1,*}, Vineet Verma¹, Xiangru Tang³, Yanjun Shao³, Eric Van Buren¹, Zhiping Weng⁴, Mark Gerstein^{5,6}, Benjamin Neale^{7,8}, Shamil R Sunyaev^{8,9}, Xihong Lin^{1,8,10}

¹ Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

² Weston High School, Weston, MA 02493, USA

³ Department of Computer Science, Yale University, New Haven, CT, USA

⁴ Program in Bioinformatics and Integrative Biology, University of Massachusetts Chan Medical School, Worcester, MA, USA

⁵ Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA

⁶ Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA

⁷ Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA

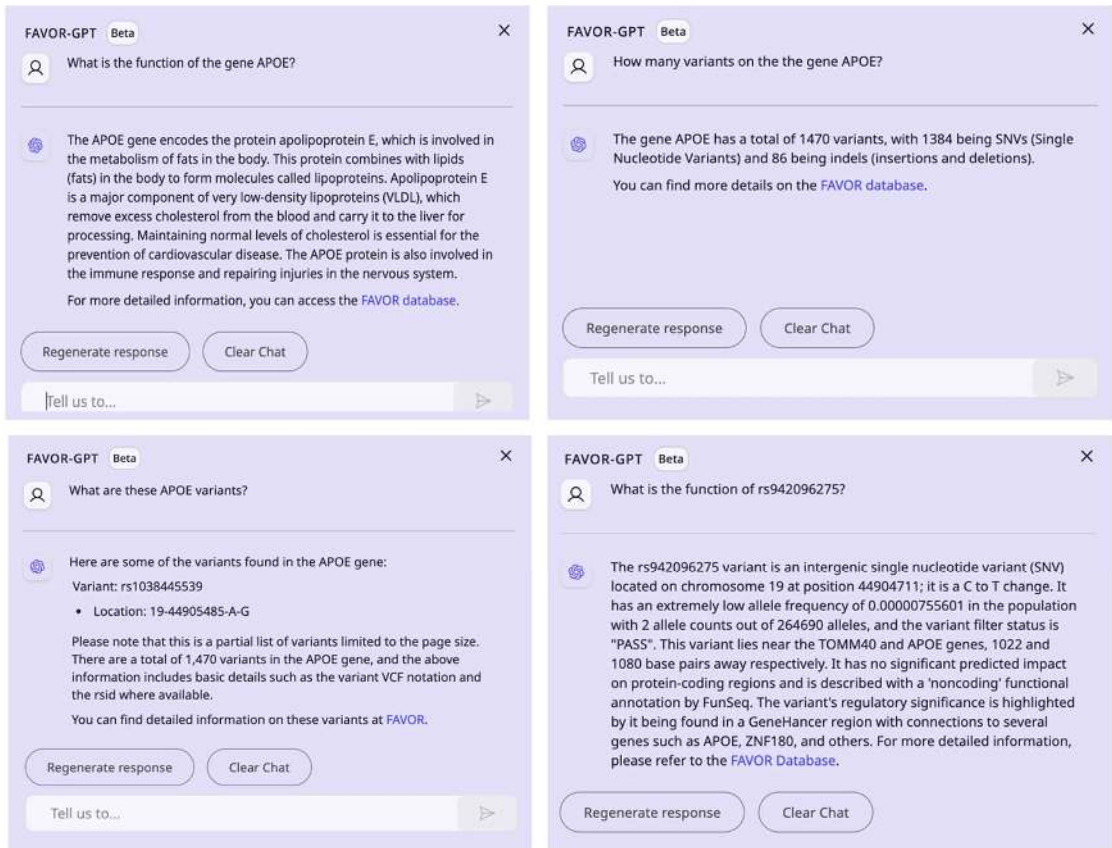
⁸ Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA

⁹ Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

¹⁰ Department of Statistics, Harvard University, Cambridge, MA, USA

* Authors contribute equally to this manuscript.

Supplementary Figure 1. Sample FAVOR-GPT responses. This example shows how FAVOR-GPT effectively interprets user input in natural language, performs precise searches in the FAVOR backend database, and then clearly explains the results in natural language.




Supplementary Figure 2. Here are some examples that querying the main web interface cannot provide but are provided by FAVOR-GPT.


FAVOR-GPT

Beta

×



How many variants of APOE have bravo_bf allele frequencies greater than 0.1



There are 6 variants of the APOE gene with Bravo_AF allele frequencies greater than 0.1:

Variant: 19-44905307-A-T

- rsid: rs449647
- bravo_an: 264690
- bravo_ac: 57264
- bravo_af: 0.216344

Variant: 19-44905910-C-G

- rsid: rs440446
- bravo_an: 264690
- bravo_ac: 181172
- bravo_af: 0.684469

Variant: 19-44907187-G-A

- rsid: rs769450
- bravo_an: 264690
- bravo_ac: 99321
- bravo_af: 0.375235

Variant: 19-44908684-T-C

- rsid: rs429358
- bravo_an: 264690
- bravo_ac: 41110
- bravo_af: 0.155314

Variant: 19-44906339-T-G

- rsid: rs1454527718
- bravo_an: 264690
- bravo_ac: 120632
- bravo_af: 0.455748


Variant: 19-44905579-T-G

- rsid: rs405509
- bravo_an: 264690
- bravo_ac: 152828
- bravo_af: 0.577385


Regenerate response

Clear Chat

FAVOR-GPT Beta ×



How many APOE variants have LINSIGHT score above 20?



There are 11 variants of the APOE gene with LINSIGHT scores above 20:


Variant: 19-44906584-C-G

- rsid: rs766970919
- bravo_an: 264690
- bravo_ac: 3
- bravo_af: 0.000011334
- linsight: 22.2466898436769


Variant: 19-44905718-C-T

- rsid: rs1479846118
- bravo_an: 264690
- bravo_ac: 1
- bravo_af: 0.000003778
- linsight: 21.6540380521622

FAVOR-GPT Beta



How many variants in APOE with aPC Epigenetics Repressed > 20



There are 367 variants in the APOE gene with an APC Epigenetics Repressed score greater than 20.

Here are some example variants:

Variant VCF: 19-44905485-A-G, **RSID:** rs1038445539, **Bravo_AN:** 264690, **Bravo_AC:** 13, **APC Epigenetics Repressed Score:** 22.2722146514031

Variant VCF: 19-44906522-G-C, **RSID:** rs1463299466, **Bravo_AN:** 264690, **Bravo_AC:** 1, **APC Epigenetics Repressed Score:** 20.4145451488832

Variant VCF: 19-44905590-G-A, **RSID:** rs1196709481, **Bravo_AN:** 264690, **Bravo_AC:** 1, **APC Epigenetics Repressed Score:** 24.2605659006957

Variant VCF: 19-44906536-T-C, **RSID:** rs1969810432, **Bravo_AN:** 264690, **Bravo_AC:** 1, **APC Epigenetics Repressed Score:** 20.8420306239427

Variant VCF: 19-44905615-G-A, **RSID:** rs1969790378, **Bravo_AN:** 264690, **Bravo AC:** 1, **APC Epigenetics Repressed Score:** 25.7279523505269

Supplementary Table 1: Sample text responses to three gene-based inputs.

Input	Output
How many variants are in SHOX	<p>The SHOX gene has a total of 238,699 variants. Within these:</p> <ul style="list-style-type: none"> • 208,805 are single nucleotide variants (SNVs) • 29,894 are insertions or deletions (indels) <p>More specific variant details, such as counts of common, low-frequency, and rare variants, are accessible through the FAVOR database. FAVOR Database SHOX.</p> <p>(Variant counts derived from Bravo Freeze 8 dataset.)</p>
What is the range for TP53 gene?	<p>The TP53 gene is located on chromosome 17 at the cytogenetic band 17p13.1. Its genomic range spans from position 7,668,420 to 7,687,489 on chromosome 17. For additional details, you can refer to the FAVOR database.</p>
Give a general description of EGFR gene.	<p>The EGFR (Epidermal Growth Factor Receptor) gene, known by its approved symbol and HGNC ID 3236, is located on chromosome 7 at the chromosomal location of 7p11.2. EGFR plays a critical role in the regulation of cell growth and differentiation. It is implicated in various cancers and is the target of several anticancer therapeutics. The gene spans from genomic position 55,019,016 to 55,211,627 on chromosome 7. The gene has other names, including ERBB1 and ERBB, and is associated with an OMIM ID of 131550. The Uniprot accession number for EGFR is P00533. It is considered an essential gene and is part of the gene family ID 1096 which includes receptor tyrosine kinases. Additionally, EGFR is highly intolerant to loss-of-function mutations, with a pLI score of 0.99968, indicative of its critical function.</p> <p>For a more in-depth look at the EGFR gene, visit the FAVOR database.</p>
How many pathogenic variants does BRCA 1 have?	<p>The BRCA1 gene has a total of 92 pathogenic variants. You can find more details on the FAVOR database.</p> <p>(Variant counts derived from Bravo Freeze 8 dataset.)</p>
How many pathogenic variants in APOE?	<p>The APOE gene has 11 pathogenic variants in total. Here is the FAVOR database link for more information.</p> <p>(Variant counts derived from Bravo Freeze 8 dataset.)</p>
How many loss of function variants in APOE?	<p>The APOE gene has a total of 12 observed loss of function (LoF) variants with an expected count of 14.714, and an observed/expected ratio of 0.81553. You can find more detailed information on APOE here.</p> <p>(Variant counts derived from Bravo Freeze 8 dataset.)</p>

How many conserved variants in APOE?	The APOE gene contains a total of 280 conserved variants. You can view more information on these variants in the FAVOR database here . (Variant counts derived from Bravo Freeze 8 dataset.)
--------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Supplementary Table 2: Comparison of the outputs generated by different models (FAVOR-GPT, RAG-T, and RAG-S and using Bravo Freeze 8 variants).

Task	Input	Model	Output
Gene Information	How many pathogenic variants are there in the gene BRCA1?	FAVOR-GPT	The BRCA1 gene has a total of 92 pathogenic variants. For further details, you can visit the FAVOR database here .
		RAG-T	There are two pathogenic variants.
		RAG-S	BRCA1 has two pathogenic variants.
	How many variants are in DMD?	FAVOR-GPT	There are a total of 577,838 variants in the DMD gene, out of which 532,367 are SNVs and 45,471 are indels. You can find more details about these variants on the FAVOR DMD gene page .
		RAG-T	There are more than 1000 variants.
		RAG-S	More than 1000
Variant Information	What is the pathogenicity of the variant 19-44908822-C-T?	FAVOR-GPT	The variant rs7412 (19-44908822-C-T) is associated with pathogenic outcomes and is clinically significant for drug response, particularly concerning Hypercholesterolemia and Familial type 3 hyperlipoproteinemia, among others. Expert panels in ClinVar have reviewed its significance and computational predictions suggest a deleterious effect on protein function. Furthermore, it influences response to medications such as Warfarin and Atorvastatin. For a comprehensive look at the variant's evidence and classifications, refer to FAVOR database .
		RAG-T	The variant 19-44908822-C-T may have clinical significance related to a specific health condition.
		RAG-S	Pathogenicity for 19-44908822-C-T needs to be evaluated with clinical correlation.
	Give a short description of rs7412	FAVOR-GPT	Rs7412 is a single nucleotide variant located on chromosome 19 at position 44908822. It is characterized by a cytosine to thymine change (C>T). This variant is found in the APOE gene and is associated with several conditions, including hypercholesterolemia, warfarin response, familial type 3 hyperlipoproteinemia, and atorvastatin response efficacy, among others. Bioinformatic predictions indicate that it is probably damaging with a PolyPhen score of 1 and a deleterious SIFT score. Its clinical significance includes associations with drug response and pathogenicity. The total allele number (AN) reported in the BRAVO database is 264,690,

			with an allele count (AC) of 20,678, and an allele frequency (AF) of approximately 7.8%. For more information, you can access the FAVOR database .
		RAG-T	rs7412 is a SNP associated with different forms of ApoE and cholesterol levels.
		RAG-S	Variant rs7412 affects the ApoE gene, linked to Alzheimer's risk and lipid profiles

AutoRecycle: Building an AI-Driven Automated Recycling Bin Using Vision Transformers

Robert Edward Owen Walmsley*
High School Student
UWC Atlantic
robbie.github@gmail.com

Thomas Cheng Li*
High School Student
Weston High School
thomaslicheng@gmail.com

Abstract

Recycling is crucial for sustainability, yet current methods are error-prone. This research introduces AutoRecycle, an intelligent recycling machine designed to automate waste sorting using advanced machine learning techniques. By leveraging the RealWaste dataset and a Vision Transformer (ViT) model trained with the Self-Supervised DINO method, our system aims to enhance sorting accuracy and reduce contamination in recycling streams. We created a proof of concept physical machine that includes a high-resolution camera, a servo motor, and a ultrasonic sensor integrated with a Raspberry Pi for real-time image processing. Experimental results demonstrate that our ViT model achieves **95.16%** accuracy in detailed waste classification and **98.74%** accuracy in binary recyclability classification. AutoRecycle proves to be a scalable prototype that can significantly improve effective recycling rates. Video demonstration and source code can be viewed at our project website: <https://robbiebusinessacc.github.io/>.

1 Introduction

A recent survey indicates that 94% of Americans support recycling, and only 35% regularly recycle, primarily due to a lack of convenient access and confusion over recyclable materials [1]. Current manual sorting methods are labor-intensive, prone to errors, and result in high costs and nearly 25% contamination in recycling streams[22]. This project aims to address these issues by developing an AI-driven automated recycling bin, AutoRecycle, which uses a Vision Transformer (ViT) model trained on the RealWaste dataset to accurately identify and sort various waste types [20] [2]. By integrating newer machine learning techniques with hardware components, this system seeks to improve sorting efficiency and provide a scalable solution for modern recycling needs.

2 Literature Review

Current automated recycling mainly uses convolutional neural networks (CNNs) and support vector machines (SVM) to separate waste with accuracies of 83% and 94.8% respectively and only separated into plastic, paper, and metal [11] [16]. Further, recycling is highly sensitive to contamination, and small amounts of wrongly recycled trash can damage the recyclability of a whole container, meaning accuracies of below 95% are not good enough for practical use [5]. Advancements in machine learning, particularly Vision Transformers (ViTs), offer a promising solution to these limitations by achieving high accuracy in image-based waste classification tasks [4]. The RealWaste dataset, with its diverse real-world waste images, is a great resource for training effective models. Unlike pristine datasets, this dataset has data from real world landfills and has enabled models like Inception V3 to reach 89.19% accuracy [18][20] [21].

*Both authors contributed equally to this work.

2.1 Vision Transformers

This paper distinguishes itself by integrating a Vision Transformer (ViT) model, trained with the Self-Supervised DINO method, into an automated recycling bin system using the RealWaste dataset [2]. Unlike traditional CNNs, which rely heavily on local feature extraction, ViTs leverage global attention mechanisms to capture long-range dependencies and contextual information more effectively. Recent literature has consistently shown the superiority of transformer-based architectures across various domains, the primary example being natural language processing models like GPT [13][14]. Transformers have also demonstrated state-of-the-art results in visual recognition tasks, where CNNs have traditionally dominated [2] [9]. Additionally, vision transformers have provided a robust framework for learning joint text-and-image embeddings such as CLIP [12], and they have helped account for how the human visual system develops from early sensory experience [10]. Because of this previous literature in this field, we were motivated to employ a ViT model, which has yielded higher accuracy compared to previous studies, including those utilizing Inception V3 [4] [23] [20].

3 Methodology

On top of creating the machine learning model, we also designed and implemented the hardware necessary to physically separate waste into recyclable and non-recyclable categories.

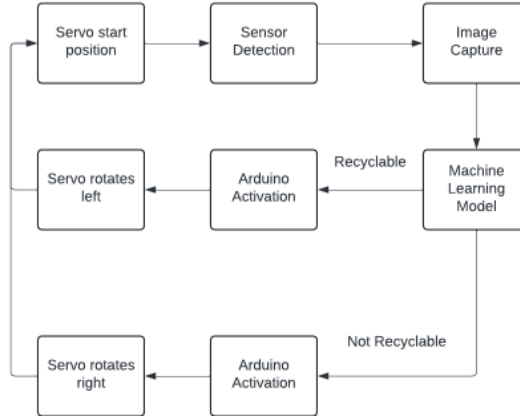


Figure 1: Diagram demonstrating the logic design of the prototype trash can

3.1 Software Design

3.1.1 Data Collection and Preparation

The RealWaste dataset comprises thousands of images categorized into cardboard, food organics, glass, metal, miscellaneous trash, paper, plastic, textile trash, and vegetation classes.

Table 1: Data Preparation Steps

Step	Description
Data Loading	The RealWaste dataset was loaded using the fastai library [6].
Preprocessing	Cleaning: Implicitly handled within the dataset loading and transformation pipeline. Resizing: Images resized to 224x224 pixels as part of the data augmentation process. Normalization: Using ImageNet statistics to match the pre-trained model’s input distribution [3]. Augmentation: Data augmentation techniques from fastai applied to create variations of training images [6].
Data Splitting	80% of the data used for training, and 20% reserved for validation.

3.1.2 Model Training

We employed a Vision Transformer (ViT) architecture for image classification tasks; specifically the ViT model trained with the Self-Supervised DINO method [2]. Although the pretrained checkpoint was self-supervised, we used supervised learning to further train it for our specific task.

Table 2: Training Process Details

Step	Description
Model Initialization	ViT model (vit_base_patch16_224.dino) initialized with pre-trained weights on the ImageNet dataset [2] [4].
Custom Head	Defined to attach to the ViT model, tailored to the number of waste categories.
Learning Rate Finder	Suggested optimal learning rates for efficient training.
Training	Initial Phase: 40 epochs with a learning rate slice between 1×10^{-3} and 1×10^{-1} . Additional Phase: 20 epochs with a lower learning rate range of 1×10^{-4} to 1×10^{-3} .
Evaluation	Performance metrics include validation accuracy, precision, recall, F1 score, confusion matrix, and top losses analysis.

3.2 Hardware Design

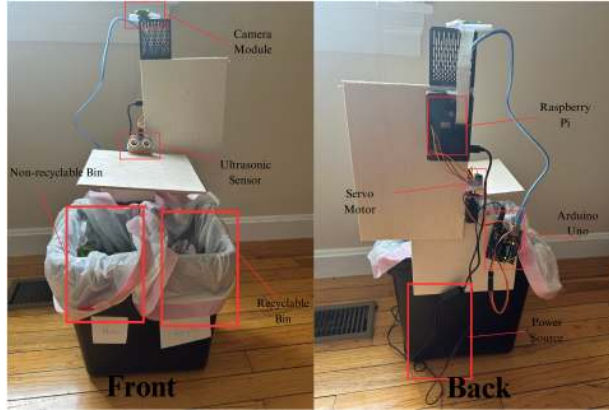


Figure 2: Labeled images of the prototype trash can (Front image on the left - Back on the right)

3.2.1 Key Components

Table 3: Key components of the AutoRecycle system

Component	Specifications	Function
Camera Module	High resolution, wide field of view	Captures waste images
Servo Motor	High torque, quick response	Directs items into categories
Ultrasonic Sensor	Accurate range and angle	Detects objects, triggers sorting
Arduino Board	Model, clock speed, I/O pins	Controls servo
Raspberry Pi	High processing power, ample memory	Runs algorithms, communicates with Arduino

3.2.2 Control Logic and Software Integration

- **Arduino Control Logic:** Programmed to receive signals from the sensor and control the servo motor. The code is written in C++ and utilizes the Arduino IDE for deployment.

- **Raspberry Pi Software Setup:** Scripts are written to get live proximity information from the sensor, capture images, process them, and send control signals to the Arduino based on the machine-learning output.

4 Results

Confusion matrix

Actual	Cardboard	97	0	0	1	0	0	1	0	0
	Food Organics	0	80	0	0	1	0	0	0	1
	Glass	0	0	85	1	0	0	2	0	0
	Metal	1	1	0	136	0	1	6	0	0
	Miscellaneous Trash	0	4	0	1	87	0	4	3	2
	Paper	0	0	0	1	0	85	0	0	0
	Plastic	1	1	0	4	5	0	175	0	0
	Textile Trash	0	0	0	0	2	0	0	61	0
	Vegetation	0	2	0	0	0	0	0	0	98
		Cardboard	Food Organics	Glass	Metal	Miscellaneous Trash	Paper	Plastic	Textile Trash	Vegetation
		Predicted								

Figure 3: Confusion matrix of the AutoRecycle system

4.1 Model Performance

The machine learning model’s performance was evaluated using a confusion matrix, top losses, and validation accuracy. Specifically, the performance analysis was divided into two key categories: detailed waste classification and binary recyclability classification. The detailed waste classification involved identifying the exact category of the waste item, such as metal, paper, cardboard, etc. The ViT model demonstrated high accuracy in this task, achieving a validation accuracy rate of **95.16%**. This indicates that the model is proficient in distinguishing between different types of recyclable materials, making it highly effective for precise sorting operations. For a more streamlined sorting process, the model was also evaluated on its ability to classify waste items into recyclable and non-recyclable categories. This is to mimic the real-life separation of single-stream recycling [7]. In this context, recyclable materials included metal, plastic, glass, paper, and cardboard, while non-recyclable materials comprised all other waste types. In this binary classification task, the model achieved an impressive validation accuracy of **98.74%**.

5 Discussion and Conclusion

Our prototype bin serves as a proof of concept for an automated single-stream recycling system. Compared to previous studies, which achieved lower classification accuracy rates of 94.8% and below, our ViT model’s **95.16%** validation accuracy in more detailed classification and **98.74%** in binary classification represent a significant advancement [19] [16]. Further, our physical prototype is a novel demonstration of how such a machine-learning model could be applied. However, limitations such as identifying a combination of different waste types remain. Future research can build on this work by expanding the dataset, adding more sensors and cameras, and scaling the system for use.

With further development, the AutoRecycle design could significantly increase recycling rates by up to 30% due to removing the fear of miss recycling from the population [8] [1]. Additionally, the decrease in recycling errors could increase the resources recycled by 20% due to decreasing contamination [22]. Overall, this intelligent recycling machine represents a promising advancement in waste management, potentially improving recycling efficiency. Recycling has huge positive impacts on climate change, water quality, and pollution. Recycling by itself could reduce carbon dioxide emissions by 6 gigatons by 2050, while reducing water pollution and air pollution from paper by 35% and 74% respectively [15] [17].

References

- [1] Survey on recycling habits in the united states, 2023. WATE.com.
- [2] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021.
- [3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
- [4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.
- [5] Ross Guberman. What is recycling contamination and how can you help? *RTS*, June 10 2020.
- [6] J. Howard and S. Gugger. Fastai: A layered api for deep learning. *Information*, 11(2), 2020.
- [7] Container Recycling Institute. Single stream recycling, 2023. Accessed: 2024-06-24.
- [8] Dhruv Kumar and Sahil Armaan Kumar. Why don't people recycle-a comparative study between the united states of america and india. *Int. J. Soc. Sci. Econ. Res.*, (2):6692–6712, 2018.
- [9] J. Maurício, I. Domingues, and J. Bernardino. Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences*, 13(9), 2023.
- [10] A. E. Orhan and B. Lake. Learning high-level visual representations from a child's perspective without strong inductive biases. *Nature machine intelligence*, 6:271–283, 2024.
- [11] M. E. Ozdemir, Z. Ali, and B. Subeshan. Applying machine learning approach in recycling. *Journal of Material Cycles and Waste Management*, 23:855–871, 2021.
- [12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, and et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.
- [13] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI blog*, 2018.
- [14] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, and et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 2019.
- [15] Celeste Robinson and Kate Huun. The impact of recycling on climate change. *Environmental Center, University of Colorado Boulder*, December 2023. Published: Dec. 15, 2023.
- [16] George E. Sakr, Maria Mokbel, Ahmad Darwich, Mia Nasr Khneisser, and Ali Hadi. Comparing deep learning and support vector machines for autonomous waste sorting. In *2016 IEEE International Multidisciplinary Conference on Engineering Technology (IMCET)*, pages 207–212, 2016.
- [17] Silila Sandawala. Paper production, recycling and climate change. *Climate Fact Checks*, January 2023. Accessed: June 25, 2024.
- [18] S. Single, S. Iranmanesh, and R. Raad. Realwaste. UCI Machine Learning Repository, 2023. DOI: <https://doi.org/10.24432/C5SS4G>.
- [19] S. Single, S. Iranmanesh, and R. Raad. Realwaste, 2023. UCI Machine Learning Repository.
- [20] S. Single, S. Iranmanesh, and R. Raad. Realwaste: A novel real-life data set for landfill waste classification using deep learning. *Information*, 14(12):633, 2023.
- [21] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.
- [22] A. Tanimoto et al. West coast contamination initiative research report, 2019.
- [23] R. Wightman. Pytorch image models. <https://github.com/huggingface/pytorch-image-models>, 2019.

Sign Language Recognition from Video using Geometrical and Transfer Learning Techniques

David Chen^{1,*,\dagger} and Thomas Li^{2,*,\dagger}

¹*Harvard-Westlake School, 3700 Coldwater Canyon Ave, Los Angeles, CA 91604, USA*

²*Weston High School, Weston, MA 02493, USA*

^{*}*Corresponding authors: dchen1561@gmail.com (DC), dbbested@gmail.com (TCL)*

^{\dagger}*Equal contribution: These authors contributed equally to this work*

August 26, 2024

Abstract

We aim to develop an American Sign Language (ASL) recognition system to bridge the communication barrier for the deaf and hard-of-hearing communities. Some previous projects utilized specialized hardware, while this study focuses on purely 2-D video stream recognition due to its accessibility. In this paper, we use the fine-tuning method, which involves fine-tuning a neural network model trained on public datasets for specific individuals in a data-efficient manner. Challenges such as image background interference and occlusion are discussed. The algorithms are tested on teenager, adult and senior male and female hands and the accuracies are comparatively better than other previous models, with the results average testing accuracy being 96.696%.

1 INTRODUCTION

The deaf and hard-of-hearing population has faced communication barriers throughout history, leading to significant challenges in their day-to-day lives. American Sign Language (ASL), a visual language used by millions of people worldwide, has become an essential means of communication for this population. However, the majority of the general population remains unable to understand ASL, exacerbating the communication divide. In recent years, machine learning and artificial intelligence have demonstrated immense potential in addressing such challenges by creating systems capable of understanding and interpreting sign language.

The primary aim of this research project is to develop and test new ASL recognition systems, capable of identifying and interpreting sign language gestures in real time. By developing a reliable ASL recognition system, this project aims to break down communication barriers for the deaf and hard-of-hearing communities, while promoting accessibility and social integration. In the long term, the successful deployment of sign language recognition systems in various applications such as education, healthcare, and public services, will contribute to creating a more inclusive and supportive society for all. While some projects have used a variety of hardware systems that normal people would not use, like special gloves, multiple cameras, etc. We wanted to test recognition systems with a single purely 2-D video stream, with a variety of background interference, lighting and hand conditions.

After reviewing the field, we have decided to employ two methods. The first is the 3-D Hand Geometry(HG) method, which uses an open source library MediaPipe to identify the 3-dimensional joint positions (landmarks) of the hand. We then created a non machine learning python code which computes the geometric relations between the landmarks, and sorts the video stream into different gestures. The second method is the Transfer Learning(TL) method, which utilizes a pre-trained neural network based on stock photo data. The model is then fine tuned based on live captured video footage of a new individual user's

40 hand, by optimizing a small portion of the neural network weights while fixing the rest. The aim of this
41 project is to compare the two methods for the purpose of detecting sign language that is effective at working
42 with all types of hands and environments.

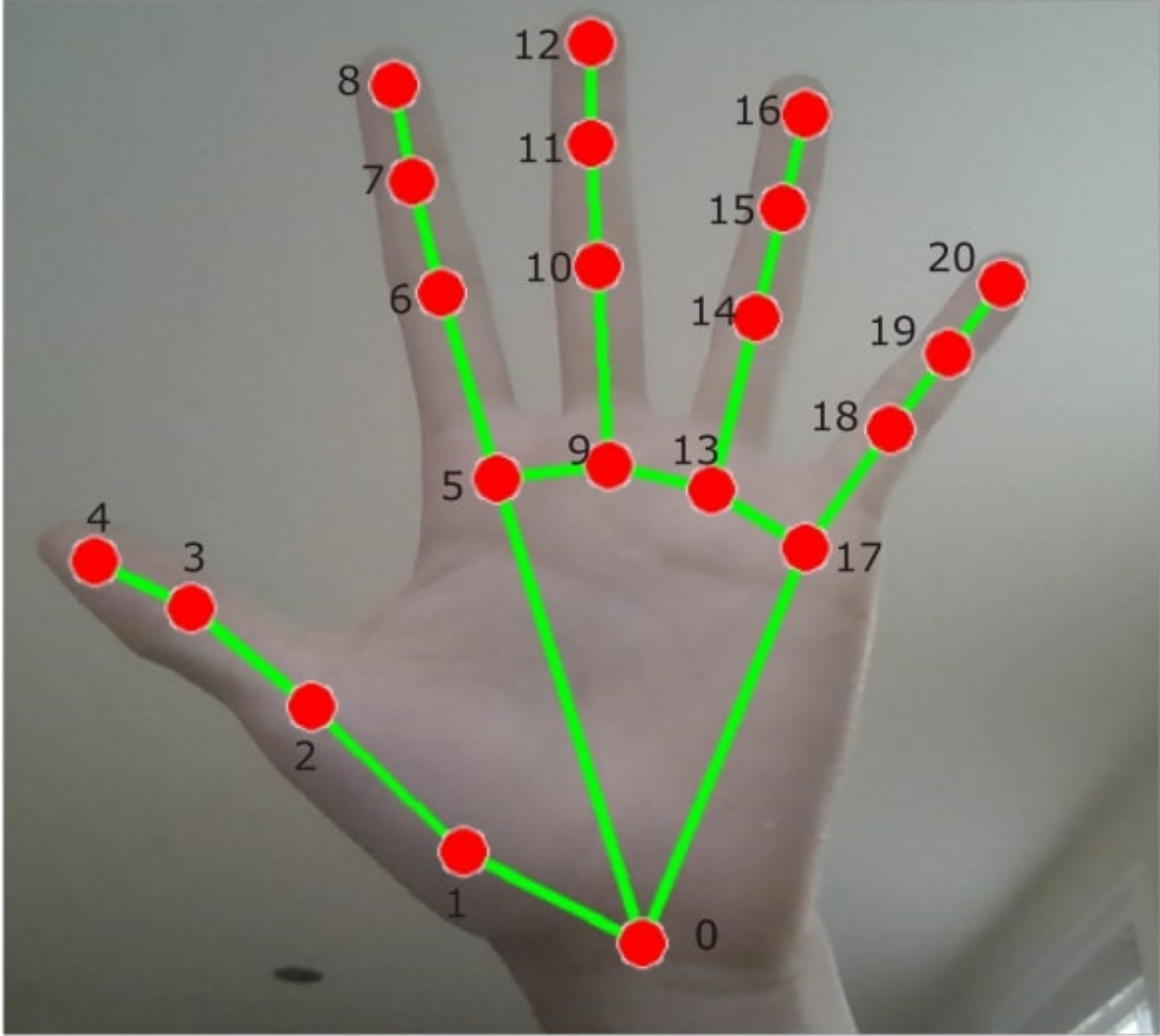


Figure 1: Visualization of Mediapipe hand landmark locations based on hand joints. Labels 1-4 are along the thumb, 5-8 are along the pointer finger, 9-12 are along the middle finger, 13-16 are along the ring finger, and 17-20 are along the pinky finger. In total there are 20 landmarks to represent the 3-D geometric state of the hand in any image.

43 2 Hand Geometry

44 In this section, we describe the Hand Geometry(HG) method, which recognizes ASL signs through the 3-
45 D geometric relations between different joints or landmarks in the hand. This method is similar to past
46 research on ASL methods, which used glove-like devices to render the hand in three dimensions and identify
47 the gestures from there.[1] In contrast to these methods, our 3-D Hand Geometry method only requires 2-D

48 pictures and videos, which makes it more applicable to general users.

49 2.1 Landmark Identification

50 The objective of the Hand Geometry method is to utilize the Mediapipe hand landmarker library [2] which
51 numerizes each finger joint into position values x, y, z . Each frame in a computer vision image can be
52 represented by its three RGB values such that $F = \{\{F_{R_i}, i = 0, \dots, n \cdot m\}, \{F_{G_i}, i = 0, \dots, n \cdot m\}\}, \{F_{B_i}, i =$
53 $0, \dots, n \cdot m\}\}$ where n and m are the dimensions of the image. The set of of hand landmarks is represented
54 as $L = \{L_i = (x_i, y_i, z_i), i = 0, \dots, 20\}$ where each (x_i, y_i, z_i) tuple is defined by the Mediapipe model M such
55 that

$$(x_i, y_i, z_i) = M(F, i) \quad (1)$$

56 where each of the xyz tuples represent the 3 dimensional location of the landmarks.

57 2.2 Landmark Geometry

58 Given the three dimensional position of all the landmarks in the hand, we then used a Python program which
59 tests all the geometric relations between each landmark in order to determine the sign output. For example,
60 the orientation of the three landmarks L_0, L_5, L_{17} give the direction of the palm depending on whether L_0 is
61 the highest landmark of the three in terms of the y coordinate. Furthermore, the relation of different joints
62 could be used to determine whether two fingers were in contact with each other or if they were inside the
63 palm area.

```

64 def palm(self, pt):
65     # form a triangle from the key points indexed by 0, 5, 17
66     self.triangle = np.array([self.xyz[0], self.xyz[5], self.xyz[17]])
67     assert len(self.triangle) == 3
68     vecs = self.triangle - pt
69     cosines = np.zeros((3))
70     cosines[0] = np.sum(vecs[0] * vecs[1]) / (np.linalg.norm(vecs[0]) * np.linalg.norm(vecs[1]))
71     cosines[1] = np.sum(vecs[0] * vecs[2]) / (np.linalg.norm(vecs[0]) * np.linalg.norm(vecs[2]))
72     cosines[2] = np.sum(vecs[1] * vecs[2]) / (np.linalg.norm(vecs[1]) * np.linalg.norm(vecs[2]))
73     count = np.sum(cosines < 0)
74     if count >= 2:
75         return True
76     else:
77         return False
78
79 ...
80
81 def letter_M(self):
82     if self.touching(self.middle_tip, self.thumb_ip, self.accuracy):
83         if self.touching(self.index_tip, self.thumb_ip, self.accuracy):
84             if self.palm(self.pinky_tip):
85                 return True
86     return False
87
88 ...
89
90 def letter_C(self):
91     if not self.palm(self.pinky_tip) and not self.palm(self.index_tip)
92     and not self.palm(self.ring_tip) and not self.palm(self.middle_tip) and
93     not self.touching(self.thumb_tip, self.middle_tip, self.accuracy):
94         if self.touching(self.pinky_tip, self.ring_tip, self.accuracy) and
95         self.touching(self.index_tip, self.middle_tip, self.accuracy):
96             if not self.palm(self.thumb_tip):

```



```

97         if self.palm_direction(self.wrist, self.index_mcp, self.pinky_mcp)[1] == "up":
98             return True
99     return False
100 ...

```

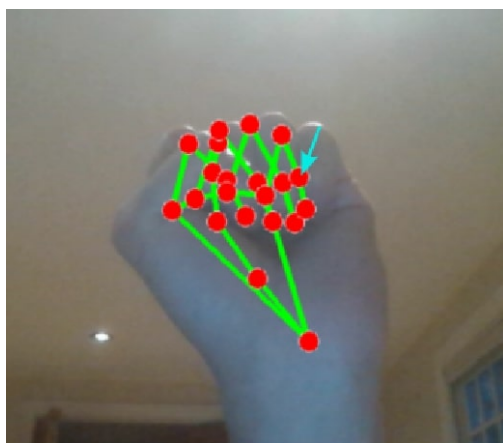
101 The above shows the Python code snippets of the 3-D HG method. A lot of "if else" statements are required
102 to be programmed and it is quite time consuming to debug the code, but we are able to finish it in reasonable
103 shape.

104 2.3 Implementation errors

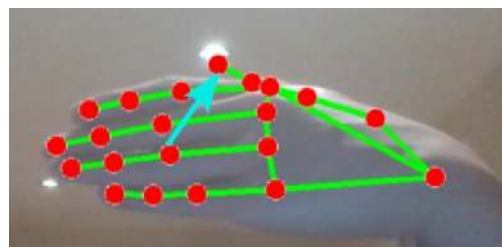
105 One of the challenges in using Mediapipe's hand landmarks for identifying sign language is the varying
106 geometric tolerances for different hands. Each person has different hand sizes, shapes and colors, leading to
107 variations in the decision statements about landmark relationships. As a result, the accuracy of identifying
108 gestures can differ between individuals. This discrepancy poses a challenge when creating a program that
109 aims to accurately recognize sign language across a diverse range of users. Adjusting the accuracy thresholds
110 to accommodate various hand variations becomes crucial but also adds complexity to the implementation.

111 Another source of inaccuracy when using Mediapipe's hand 3-D landmarks for sign language identification
112 is the image background interference that can affect landmark detection. The hand tracking algorithm relies
113 on distinguishing the hand from the background, and any elements in the environment that resemble or
114 overlap with the hand can interfere with accurate landmark labeling. For instance, if the background for the
115 picture or video is a similar color to the user's hands, it will cause identification errors where the background
116 becomes identified as a hand. These background factors can introduce errors in recognizing sign language
117 gestures, compromising the overall accuracy of the system.

118 In sign language, different joints and regions of the hand are utilized to form specific signs or gestures.
119 However, when performing certain signs, it is possible that some hand joints or landmarks get occluded
120 or blocked from the camera's view. This occlusion can occur when the hand crosses over itself, when
121 fingers overlap, or when certain hand configurations obscure specific landmarks. As a result, the missing
122 or obscured landmarks can lead to inaccuracies in identifying the intended sign language gesture. Dealing
123 with occlusion scenarios becomes a significant challenge in leveraging Mediapipe's hand landmarks for sign
124 language recognition, as it requires additional techniques or algorithms to handle partial or incomplete
125 landmark information.



(a) Error Example 1



(b) Error Example 2

Figure 2: Error Example 1 shows landmark errors when joints are close together. Error Example 2 shows landmark errors with background objects identified as joints.

3 Transfer Learning Method

In this section, we describe the second methodology of recognizing ASL signs from 2-D RGB video streams. We start with briefing the problem definition of ASL alphabet recognition in Section 3.1. Next we introduce the fine tuning technique applied in Section 3.3, which is a common technique in neural network based deep learning. In Section 3.4, we present our software design on mobile device that assists few shots data collections on a new domain (i.e. the hand of a new user) for the purpose of model fine tuning. Related implementation details are provided in Section 3.5.

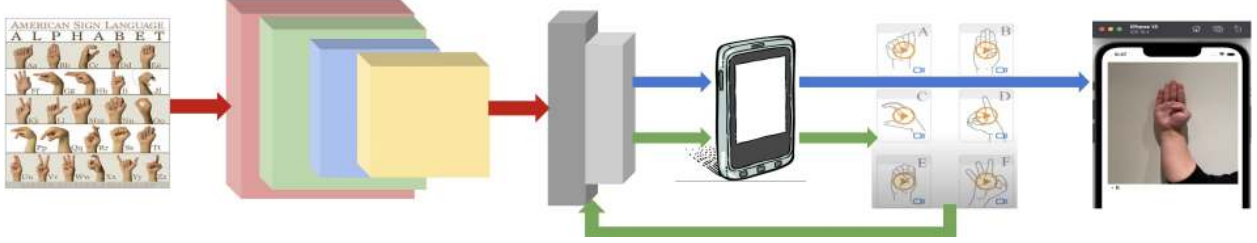


Figure 3: The proposed ASL recognition system featuring neural network model fine tuning on mobile device. **Red arrow flow**: offline ASL model training pipeline using public datasets. **Green arrow flow**: ASL model fine tuning using individualized collected datasets. **Blue arrow flow**: fine tuned ASL model online inference on mobile end. Our ASL recognition model takes more than 85,000 video frames and predicts ASL letters given mobile captured video stream in 30 fps.

3.1 Problem Definition

The objective is to determine a single ASL letter out of 26 alphabets and 3 special characters [3] from streamed video frames. Because some of the signatures in the alphabet such as "Z" and "J" are motion based, the frames in computer vision's terminology are represented as a sequence X of RGB images in terms of $X = \{x_{RGB,t}\}_{t=1}^T$, where the length of video frames is denoted by T . The set of ASL alphabets is represented as $Y = \{y_i, i = 1, \dots, 29\}$. Our ASL recognition problem hence can be defined as learning a model \mathcal{M} such that

$$y_t = \mathcal{M}(x_{RGB,t}) \quad (2)$$

where $y_t \in Y$ given the timestamp t is the recognized ASL letter. Because the prediction of the highest likelihood letter is the optimal one from 29 ASL characters, this problem is considered a classification problem, and therefore \mathcal{M} would be a classification model with neural network weights to be optimized. For all timestamps $t = 1, \dots, T$, $\mathcal{M}(x)$ would represent the model output and $y(x)$ would be the vector representation of the label, with the vector value being 1 at the index where the classification is correct and 0 everywhere else. Therefore if $\mathcal{M}(x)$ has the highest value at the same index where $y(x)$ is equal to 1, and furthermore if this highest value is very positive, the model would be more accurate for that one instance.

3.2 Model Learning

Given the classification problem stated in Eq. 2, we apply a similar modeling approach with traditional image classification deep neural networks [4] in order to address object classification on image domain. The model we trained can be decomposed into the following:

$$\mathcal{M} = \{\mathcal{M}_{BB}, \mathcal{M}_{FC}\} \quad (3)$$

where \mathcal{M}_{BB} represents the backbone network layers (BB) that encode the RGB images to multi dimensional features. \mathcal{M}_{FC} is a fully connected layer (FC) placed at the end of \mathcal{M} . It maps the encoded features to the

layer of classification output. The objective is to minimize the so-called cross-entropy loss[5] between the classification predictions and target:

$$L(x, y) = \frac{\sum_{n=1}^N l_n}{N} \quad (4)$$

where

$$l_n = - \sum_{c=1}^C \omega_c \log \frac{\exp(\mathcal{M}(x)_{n,c})}{\sum_{i=1}^C \exp(\mathcal{M}(x)_{n,i})} y(x)_{n,c} \quad (5)$$

where $\frac{\exp(\mathcal{M}(x)_{n,c})}{\sum_{i=1}^C \exp(\mathcal{M}(x)_{n,i})}$ is the probability predicted by the model of the label being in class c , and $y(x)_{n,c}$ is either 0 or 1 depending on whether the label is of class c . ω_c is the individual weight for each class, which is used because of the different number of appearances for each class in the training dataset. Because ω_c is there to counteract the training set inequality, it is a fixed value and does not change during model learning. As the model at the moment is a classification model for classifying between 29 classes, C is 29.[5]

In this work we used different optimizers, including RMSProp(RMSP), Adam, and AdamW, to obtain the optimal neural network parameters for $\mathcal{M}(x)$ that minimizes cross entropy loss.

3.3 Model Fine Tuning

Following the learning objective defined in Eq. 4, model \mathcal{M} can be trained by iterating through the dataset points $\{x, y(x)\}_D$. Evaluation of the model can be performed on the subset $D_{sub} \in \{x, y(x)\}_D$ which is unseen during the training. In real-world applications, ASL video frames are collected *in-the-wild* with diverse backgrounds and hand size, shape and color profiles. Those frames may fail at classification for initial \mathcal{M} because the image data, in appearance, can be quite different from the dataset applied to model training. And as we can see in Sec. 2, these background factors can wildly affect the output results in a negative way.

To attain a more applicable model, we apply the fine tuning technique [6]. The core strategy of fine tuning a neural network is to continue the training referring to Eq. 4 on a fixed set of layers in model \mathcal{M} while keeping the rest of layers weights unchanged. Specifically, the model weights of \mathcal{M} are categorized as:

$$\{\omega, \omega \in \mathcal{M}\} = \{\omega_{fx}, \omega_{fx} \in \mathcal{M}_{BB}\} \cup \{\omega_{ft}, \omega_{ft} \in \mathcal{M}_{FC}\} \quad (6)$$

where ω_{fx} are backbone layers in which the weights are fixed for feature extraction while ω_{ft} are weights of the final fully connected layer updated for fine tuning. The weights update of fine tuning layers \mathcal{M}_{fc} follows the learning objective defined in Eq. 4 and 5 as to solve the same classification problem.

3.4 Mobile Application

The approach described in Sec. 3.3 is suitable for use cases where the new image frames are unseen to the base model but are adapted through fine tuning to achieve the same learning objective. To assist with getting a more user-friendly interface of our proposed method, we take advantage of an Apple iPhone IOS based app *Translitero* and port the pre-trained model weight $\{\omega_{fx}, \omega_{ft}\}$ based on public hand-gesture dataset to the mobile end in Fig. 3, to provide the portability. The user interface allows the data collection process of the new use cases to be natural and fast.

In Fig. 3 we sketch the core components that supports the fine tuning in a data-efficient manner:

- **Data Collection.** The user interface opens up the camera to collect live video stream of the new user's hand gesture for up to ten seconds. The user is prompted to make hand gestures of the 29 ASL letters, and the recorded video clips are automatically labeled.
- **Fine Tune Backend.** The collected data in video clips(MP4) are uploaded to a desktop computer server where a pipeline takes the RGB video frames and then follows the procedure detailed in Sec. 3.3 to optimize and obtain the fine tuned neural network weights. The fine-tuned neural network weights ω_{ft}^{new} are then sent back to the mobile phone.

- **Live Inference.** The real-time video stream captured from mobile camera is projected to the handheld screen. In the meantime, the ASL model $\{\omega_{fx}, \omega_{ft}^{new}\}$ takes the raw frames and predicts the most likely ASL letter based on model inferences directly on the smartphone.

The session time of data collection is set as one minute maximum per ASL letter. The collected video clip from iPhone device has the frame rate of 30Hz. As demonstrated in section 4, 10 seconds recording (~ 300 frames) per ASL letter applied to model fine tune could attain an acceptable recognition accuracy, thus requiring less than 5 minutes recording time in total. It justifies our proposed approach in realizing in-the-wild ASL recognition as well as the mobile user interface that provides a user-friendly and rapid data collection method.

3.5 Implementation Details

The fine-tuning method of ASL letter classification, proposed in Sec. 3.2, is implemented under PyTorch deep learning code framework [7]. The architectural details are below:

- **Back Bone.** Following the definition of \mathcal{M}_{BB} in Equ. 3, we adopt MobileNetV2 [8] as the back bone architecture with pretrained weights that takes RGB image frames in size of 224 x 224.
- **Classification.** Following the definition of \mathcal{M}_{FC} in Equ. 3, the classification layer is a fully connected (FC) one with dimension of 29 corresponds to the size of ASL letters introduced in this work.

For model training on \mathcal{M} , we apply Adam optimizer [9] with the learning rate of 0.0005, and other optimizers. A learning rate scheduler is also utilized during the training. The same optimization process is also adapted for the model fine tuning.

Translitero integrates multiple view controllers as the user interfaces developed in XCode 14.0 and IOS 14.5 to satisfy data collection or online ASL inference. Portability of the ASL model is supported by LibTorch 10.0 framework installed on an iPhone.

Methods	AA (%)	TM 5s (%)	TM 10s (%)	TM 20s (%)
HG	78.60	85.50	85.44	85.43
MASL	97.76	20.03	20.00	19.99
MASL+FT (Proposed)	N/A	90.76	95.45	95.89

Table 1: ASL letter recognition accuracy in percentages. **AA**: results on ASL Alphabet dataset. **TM**: results on Translitero Mobile dataset. 5/10/15 seconds correspond to frame time length used for model fine tuning. The model fine tuning approach we proposed demonstrates the better accuracy.

Optimizers	RMSP [10] (10s) (%)	Adam [9] (10s) (%)	AdamW [11] (10s) (%)
Accuracy	95.33	95.45	95.48
Blurred Background	MASL + FT (5s)	MASL + FT (10s)	MASL + FT (20s)
Accuracy	89.99	95.20	95.56

Table 2: Ablation study of fine tuning based ASL letter recognition accuracy in percentage. **Row 1-2**: results of adapting different optimizers on model fine tuning under the same data collection time (10s); **Row 3-4**: results of applying image frames with blurred background but different data collection time during model fine tuning.

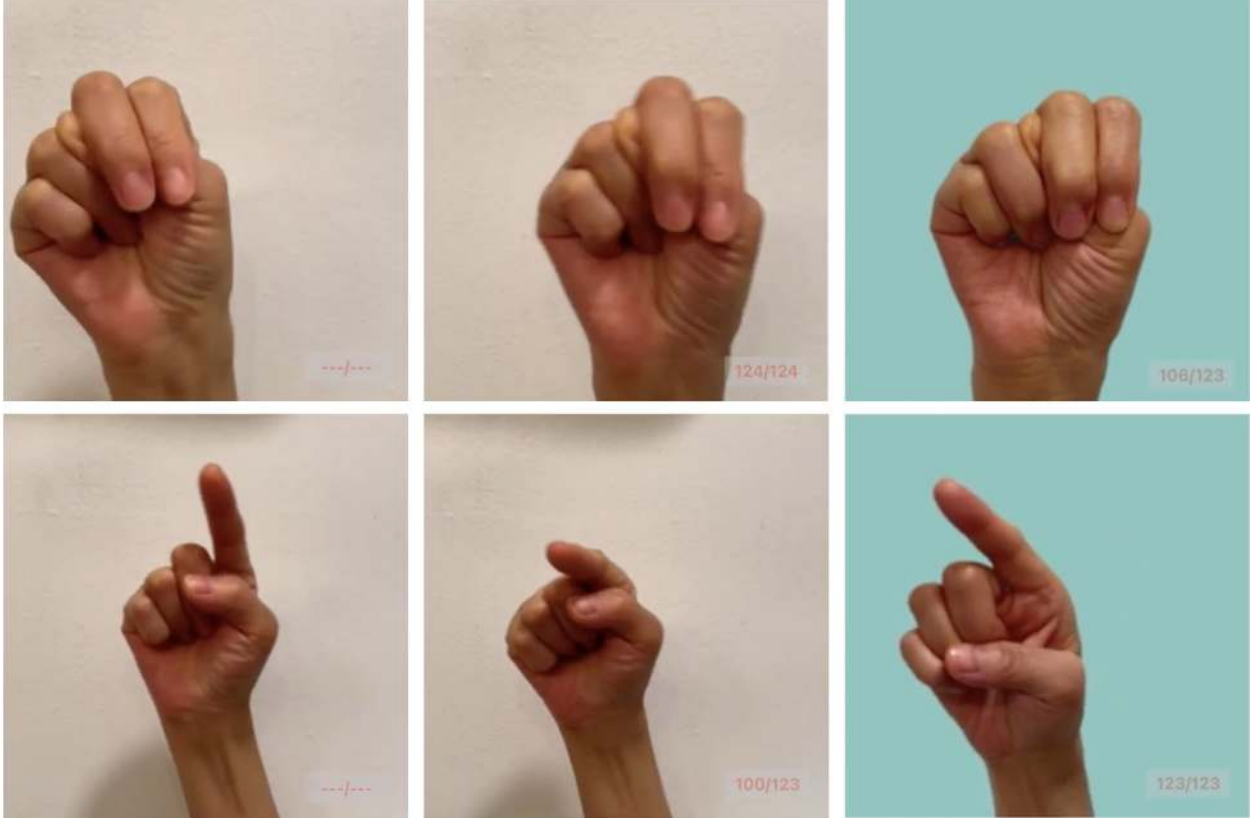


Figure 4: Visualization of ASL letter recognition qualitative examples. **Column 1**: ASL gesture frame samples captured from mobile camera using the app. **Column 2**: ASL letter recognition accuracy (bottom-right red fonts) tested on pre-recorded ASL clips. **Column 3**: verification on recognition accuracy while applying virtual background to the clips. **Row 1** and **Row 2** correspond to ASL letters **N** and **Z**.

4 Experiments

We have carried out ASL recognition experiments on teenager, adult and senior male and female hands in the wild and the accuracies obtained are compared.

4.1 Datasets

- **ASL Alphabet.** The training data contains 87,000 images in 29 ASL letter categories, are 26 letters A-Z and 3 classes for SPACE, DELETE and NOTHING, which are necessary for spelling words and sentences.
- **Translitero Mobile.** Short-duration ASL alphabets video frames are captured using an iPhone. Those video clips are 1 minute in length containing 29 ASL letter categories to validate the model's ability of generalization.

The ASL Alphabet dataset is further divided into training and testing splits for model learning and evaluation purposes. Video segments from the Translitero mobilephone-collected dataset are sampled in fixed length for model fine tuning and transfer learning. The rest of the video segments are used for evaluation. Variation of the sample video frame length used for fine tuning is discussed in Sec. 4.3

4.2 Methods

Among the various methods tested, the most effective methods for ASL letter recognition are:

- **Hand Geometry(HG).** This method extracts 3-D hand key points (landmarks) based on video image frames utilizing the Mediapipe [12] library. The geometry-based classifier on 29 ASL letters are derived based on hand key points geometric relations.
- **MobileNet ASL(MASL).** The vanilla ASL classification model developed according to Sec. 3.2, 3.5 using public datasets with no customization. The model is trained end-to-end from image frames to 29 ASL classes upon ASL Alphabet dataset.
- **Fine Tuned MASL(MASL+FT).** The ASL classification model which uses the fine tuning technique (Sec. 3.3) using new individualized data from Translitero Mobile clip frames.

4.3 Results

4.3.1 Quantitative Results

We summarize the quantitative results in Table 2 of the recognition accuracies evaluated on ASL Alphabet and Translitero datasets. The percentages are averaged over the 26 ASL letters A-Z. HG accuracies are all below 90% which indicates that the 3-D geometric cues extracted from hand key points are not precise enough to attain exceptional recognition accuracy. Results of MobileNet ASL(MASL) trained on ASL Alphabet dataset demonstrates the great performance if tested as well on the ASL Alphabet public dataset. However, the low recognition accuracies on Translitero Mobile dataset collected on new individuals indicates the model does not have the ability of generalization. During test time, a few of the ASL letters recognition accuracies are observed to be nearly zero. On the other hand, after employing the model fine tuning technique to MobileNet ASL model(MASL+FT), the issue of generalization gets largely mitigated. The 10s video frame length per ASL letter applied for fine tuning could achieve the test accuracy above 95%. Those quantitative results validate our transfer learning approach on improving performance of neural network models with high data efficiency.

4.3.2 Qualitative Results

In Fig 4, we selected two ASL letter recognition cases, "N" (row 1) and "Z" (row 2). The selected two cases are challenging according to our observations. Letter *N* is similar to *M* in appearance, with the only difference being in the position of the thumb tip relative to the fist. *Z* is a dynamic letter, which means that it involves the hand moving during the duration of the gesture. Since our modeling approach predicts per frame, there is a chance that the dynamic letters could suffer from lack of data. Despite that, most of our testing results showed 100% accuracies, which proves the quality when executing the fine tuned model.

4.3.3 Ablation Study

In machine learning, an ablation study is when we study the effects of parts of the model by removing or silencing certain layers. We conducted this type of study upon our fine tuned model to explore more alternatives that could obtain the optimal performance in terms of ASL letter recognition accuracy. In this work, we investigate the following variants:

- **Optimization Algorithm.** During the model fine tuning phase, the weights of fine tune layers, ω_{ft} , are updated according to the gradient descent algorithm. Choices of optimization could result in different efficiency and accuracy.
- **Virtual Background.** Considering the high background diversity of the video image frames captured on the mobile device, we additionally develop a functionality that enables attaching a customized background to the video stream. Inside the camera view, the background is blurred resulting in only

the moving hands being visible. We apply this setup on both the fine-tuning data collection stage and real-time ASL inference stage.

In Table 2, among the listed optimization algorithms, the Adam optimizers [9, 11] attains better accuracy than RMSP [10]. And AdamW [11] could achieve the best accuracy given 10 seconds of fine-tuning data collection. Further, we observe that the ASL recognition accuracy is not having much performance change with various physical backgrounds (curtains, lighting, wallpaper) after applying the virtual background. For all the fine tuning collection time tested on 5, 10, 20 seconds, the accuracy results are close to the uniform background ones. Those are also demonstrated in the qualitative results in Fig. 4. These results prove the effectiveness of developing the virtual background feature in our mobile APP such that the impact of physical backgrounds to the ASL model could be mitigated to greatly enhance the usability.

Letter/ Demo-graphic	Teenage Male	Teenage Fe-male	Adult Male	Adult Female	Senior Male
C	100%	73.2%	100%	100%	100%
E	100%	100%	100%	100%	100%
H	89%	100%	100%	100%	100%
O	100%	89.4%	100%	100%	100%
R	94.2%	91.1%	100%	100%	63%
SPACE	100%	100%	100%	100%	100%

Table 3: The above shows the hand gesture data collected from the mobile app. ASL finger spelling recognition was analyzed to assess the accuracy across categories of a teenage male, a teenage female, an adult male, an adult female, and a senior male. The accuracy is determined by the correct number of hand gesture predictions divided by the total frames, with the number of frame ranging from 80 to 123 depending on the device. The recognition system achieved high accuracy for teenage males and females as well as adult males and females. However, the accuracy is still less satisfactory for a senior male, likely due to factors such as hand tremors and variations in hand shape or size.

5 CONCLUSIONS

In conclusion, our research translates ASL to English with greater than 95% accuracy using only 2D video image-based inputs. The transfer learning method proved to be highly data-efficient during the training stage and was a lot more effective during testing due to its adaptiveness to different hands and environments. Furthermore, the mobile app with the transfer learning algorithm can be used in practical circumstances with different physical backgrounds that occur in real life. This could mean that ASL learning for the sign language community could be much easier in the future.

The ASL algorithms and app may be further expanded into more areas. The app can be made to accommodate more than just the ASL alphabet, such as complete words. In addition, the sign language transfer learning method can be applied to more sign language systems such as the Spanish / French / Russian Sign Languages.

Acknowledgement

We acknowledge the patient teaching and guidance by Mr. Yingchun Chen and Mr. James Ma.

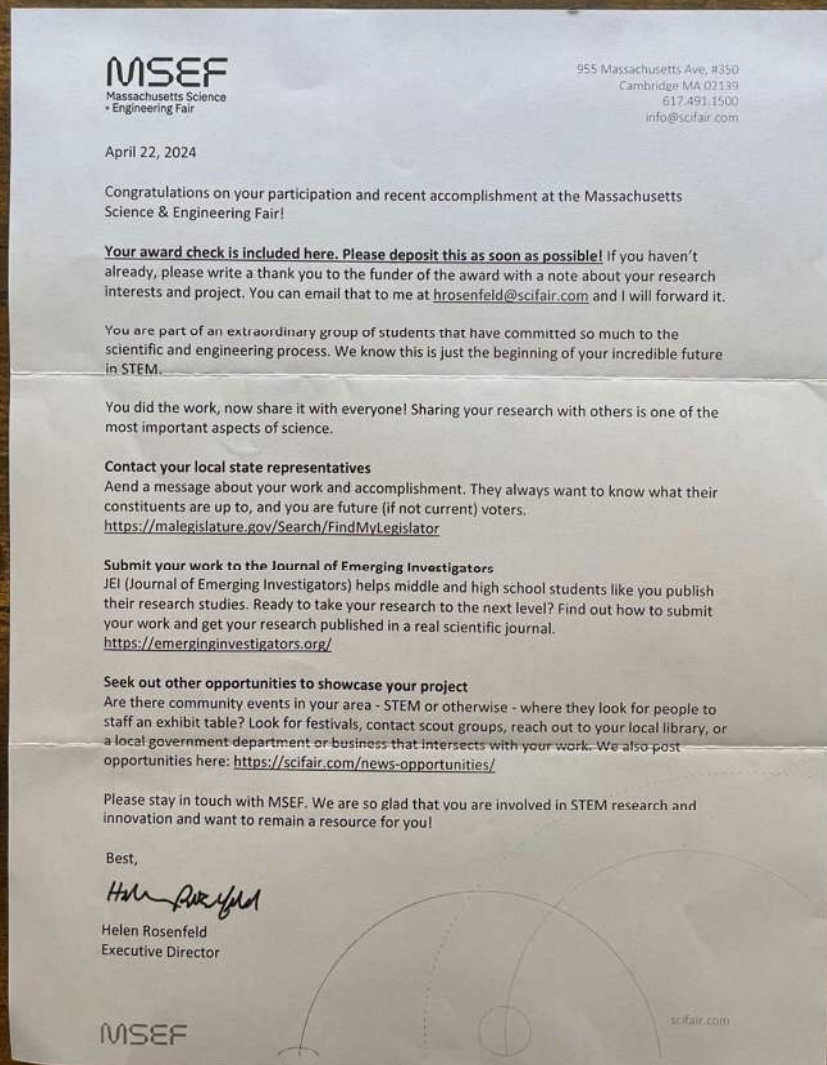
References

- [1] Zhihao Zhou, Kyle Chen, Xiaoshi Li, Songlin Zhang, Yufen Wu, Yihao Zhou, Keyu Meng, Chenchen Sun, Qiang He, Wenjing Fan, Endong Fan, Zhiwei Lin, Xulong Tan, Weili Deng, Jin Yang, and Jun

- 296 Chen. Sign-to-speech translation using machine-learning-assisted stretchable sensor arrays. *Nature*
297 *Electronics*, 3:571–578, 2020.
- 298 [2] Google. Hand landmarks detection guide. [https://developers.google.com/mediapipe/solutions/](https://developers.google.com/mediapipe/solutions/vision/hand_landmarker)
299 [vision/hand_landmarker](https://developers.google.com/mediapipe/solutions/vision/hand_landmarker).
- 300 [3] Asl alphabet. <https://www.kaggle.com/dsv/29550>.
- 301 [4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional
302 neural networks. *Communications of the ACM*, 60(6):84–90, 2017.
- 303 [5] Pytorch. Crossentropyloss. [https://pytorch.org/docs/stable/generated/torch.nn.](https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html)
304 [CrossEntropyLoss.html](https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html).
- 305 [6] Joanne Quinn, Joanne McEachen, Michael Fullan, Mag Gardner, and Max Drummy. *Dive into deep*
306 *learning: Tools for engagement*. Corwin Press, 2019.
- 307 [7] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming
308 Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.
- 309 [8] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mo-
310 bilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer*
311 *vision and pattern recognition*, pages 4510–4520, 2018.
- 312 [9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint*
313 *arXiv:1412.6980*, 2014.

- 314 [10] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture
315 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2, 2012.
- 316 [11] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint*
317 *arXiv:1711.05101*, 2017.
- 318 [12] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan
319 Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building
320 perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

3rd Place Award
at the Massachusetts Science & Engineering Fair (MSEF) 2024
*"Analysis and Machine Learning Modeling of Spatial Data to Identify
Asthma Hotspots in Massachusetts"*



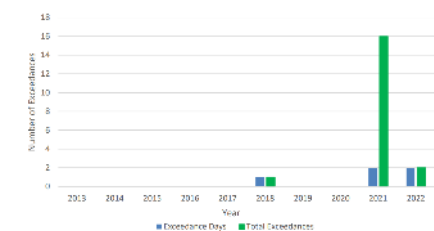
Analysis and Machine Learning Modeling of Spatial Data to Identify Asthma Hotspots in Massachusetts

Thomas Li, Weston High School

BACKGROUND

Asthma's uneven distribution across Massachusetts demonstrates a pressing public health issue, with specific areas experiencing disproportionately high rates of the condition. This disparity is often attributed to a complex interplay of environmental and socio-economic factors, underscoring the need for a detailed spatial analysis to identify asthma hotspots. This research is crucial for understanding the geographical nuances of asthma incidence, facilitating targeted interventions, and addressing environmental injustices that contribute to health disparities. Further, the rapidly evolving environment will mean that the current Asthma medical data will become quickly outdated.

Figure 3
24-hour PM_{2.5} Exceedance Days and Total Exceedances 2013-2022
Exceedances Based on 35 µg/m³ 24-hour Standard



2.5-micrometer Particulate Matter concentration exceedances per year. Demonstrates the unpredictable nature of environmental factors.
Massachusetts 2022 Air Quality Report(<https://www.mass.gov/doc/2022-annual-air-quality-report/download>)

QUESTION AND HYPOTHESIS

Question: Can machine learning models be used to predictively generate the areas of Massachusetts that are most prone to asthma?

Hypothesis: A Random Forest machine learning model can classify geographical locations as asthma hot spots with high accuracy based on environmental data.

REVIEW OF LITERATURE

Machine learning models have been extensively used to predict asthma exacerbations with various degrees of success. These models incorporate a range of predictors including demographic, clinical, and socioeconomic factors. The most commonly used machine learning algorithms are logistic regression and random forests, but other techniques like XGBoost and LightGBM are also in use. The number of predictors in these models can range widely from 1 to over 200, and they often include systemic steroids use, beta2-agonists, emergency department visits, age, and previous asthma exacerbation history. The performance of these models is quite variable, with the area under the receiver operating characteristic curve (AUROC) ranging from 0.59 to 0.90, indicating that while some models are excellent, others may have limited predictive power.

When focusing specifically on predictive models for asthma attacks, the use of biosignals and environmental factors is common. The majority of studies concentrate on biosignal risk factors, but a substantial number also factor in environmental triggers. The data acquisition methods for these models vary, including telemonitoring technologies, emergency department records, national databases, and environmental data from meteorological agencies and pollution monitoring stations. However, it's worth noting that having a large and varied population size is crucial for the reliability of the study's findings. The generalizability of these models is a challenge that requires further research across different groups and populations.

METHODOLOGY

1) Aggregate data

I collected environmental data such as particulate matter concentration from MassDEP and pediatric asthma data from Massachusetts Environmental Public Health Tracking.



Location of air monitoring stations as set up by the Massachusetts Department of Environmental Protection.
Source: Massachusetts 2021 Air Quality Report(<https://www.mass.gov/doc/2021-annual-air-quality-report/download>)

3) Build training set

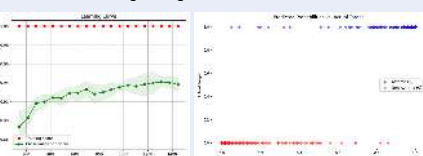
Randomly select spatial data points concentrated so that the areas with higher prevalence rates are equally represented



Points from a random sampling of high asthma concentration areas and low asthma concentration areas.
Made in ArcGIS Pro by student.

4) Train ML model

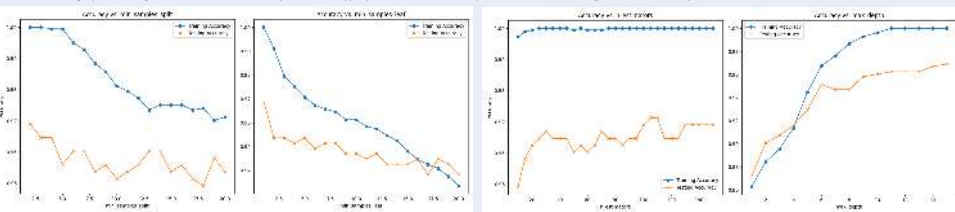
Train models using a variety of different hyperparameters and observe the resulting testing accuracies.



Model training learning curves. Displays the training and cross validations scores as well as the predicted probabilities versus actual target.

5) Tune hyperparameters

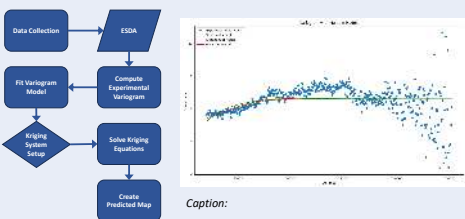
Experimented with hyperparameters min sample split, min sample leaf, n estimators, and max depth. Min Sample Split and Min Sample Leaf prevent overfitting by controlling node splitting and leaf size, ensuring trees don't learn the training data too precisely. N Estimators determines the forest's size, affecting accuracy and computation time. Max Depth limits tree complexity, helping avoid overfitting by restricting the number of decision layers. These hyperparameters collectively fine-tune the model's generalization and performance.



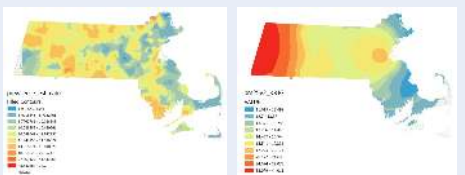
Final training versus testing accuracies with the hyperparameters as the independent variable and the accuracy as the dependent variable. All other parameters are held constant.

2) Interpolate spatial data

Data collected using monitoring stations is only for a specific location. Therefore, we have to interpolate the monitored air quality onto the surrounding area. We use the **Kriging interpolation** method in this instance, we use the Kriging interpolation method because of its ability to incorporate the spatial autocorrelation of the sampled data effectively. This method is particularly suitable for air quality data, which often exhibits strong spatial correlation. Kriging interpolation is a method that predicts values for unsampled points by utilizing the spatial correlation of the sampled data, as quantified by semivariance.



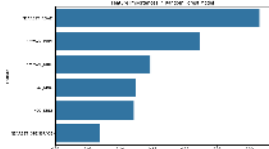
The variogram above is calculated from spatial data, showing how data similarity decreases with distance. Each point represents the semivariance between pairs of data points at a specific lag distance, revealing the spatial structure and continuity within the dataset. The pattern depicted by the points is used to model the variogram, which is fundamental for the Kriging interpolation process. This model helps in understanding the spatial correlation and is crucial for predicting values at unsampled locations with greater accuracy.



Resulting interpolation maps for Asthma prevalence and PM 2.5 levels displayed in ArcGIS Pro and made by student.

RESULTS

- The Random Forest model was able to predict whether a specific location was an asthma hotspot with **96.9% accuracy** and an **F1-score of 97.5%**.
- The highest accuracies were observed in models with 110 estimators, 15 max depth, 2.5 minimum sample split, and 2.5 minimum sample leaf.
- Of the various features tested, **distance to the nearest road** was the most important factor, with **PM2.5 levels** the next most important



CONCLUSIONS

- We created a Random Forest model to predict **pediatric asthma hotspots in Massachusetts**, using a variety of publicly-available data.
- The final model was able predict whether a specific location was an asthma hotspot with **96.9% accuracy**.
- The model suggests that the most critical Environmental factors for identifying geographical asthma hotspots are **distance to roads** and **particulate matter concentrations**, especially PM2.5.
- The areas with the highest predicted asthma risk are the regions near **Springfield and Greater Boston**.
- The final interpolated map of predicted asthma hotspots is below:

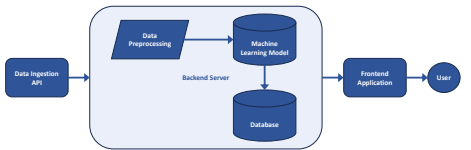


- This model can be used to help public policy experts, families, and especially children **identify areas of high asthma risk**, with the goal of aiding prevention and treatment efforts.

FUTURE WORK

Areas of further research could include:

- Predicting asthma hotspots for the current year.
- Building a website or app that displays the hotspots in real time, so the results can be more easily accessible to the public.



- Enriching the training dataset using additional features (such as ambient air temperature, average household income, demographic information, etc.), as well as additional historical data.
- Training additional models and fine-tuning to further improve results.

KEY REFERENCES

- Wang, M., Qiu, J., & Ong, P. (2018). Machine Learning in Spatial Study of Asthma Rate Distribution in Los Angeles County. 2018 Chinese Automation Congress (CAC), 3696-3700. <https://doi.org/10.1109/CAC.2018.8623677>.
- Razavi-Termeh, S.V., Sadeghi-Niaraki, A., & Choi, S.M. Asthma-prone areas modeling using a machine learning model. Sci Rep 11, 1912 (2021). <https://doi.org/10.1038/s41598-021-81147-1>
- Luo, G., Johnson, M., Nkoy, F., He, S., & Stone, B. (2020). Automatically Explaining Machine Learning Prediction Results on Asthma Hospital Visits in Patients With Asthma: Secondary Analysis. JMIR Medical Informatics, 8. <https://doi.org/10.2196/21965>.
- Emanet, N., Öz, H., Bayram, N., & Delen, D. (2014). A comparative analysis of machine learning methods for classification type decision problems in healthcare. Decision Analytics, 1, 1-20. <https://doi.org/10.1186/2193-8636-1-6>.



Thomas Li <25lit@weston.org>

2024 Simons Summer Research Program at Stony Brook University - Thomas Li

Simons SRP <simonssrp@stonybrook.edu>

Tue, Mar 19, 2024 at 6:15 PM

To: Thomas Li <25lit@weston.org>

Cc: Alisa Yurovsky <alisa.yurovsky@stonybrook.edu>, Alisa.Yurovsky@stonybrookmedicine.edu, Karen Kernan <karen.kernan@stonybrook.edu>

Thomas,

I am delighted to inform you that you have been admitted to the **2024 Simons Summer Research Program** based on the quality of your application, your strong academic record and the recommendations of your teachers. The Simons applicants include some of the best science students from all over the country. Your nomination by your high school for this fellowship is in itself an honor and an indication of your exceptional promise. And we have no doubt that you will make the most of this opportunity.

The Simons Summer Research Fellowship Program runs from **July 1st through August 9th, 2024**. During this time you will be working under the mentorship of Dr. **Alisa Yurovsky** in the Department of Biomedical Informatics. In addition to your research, you will be expected to attend all Simons Program activities, such as weekly lunches/informal research talks; and will be presenting at the closing poster symposium on August 9th. Upon successful completion of the program, all participants will be awarded a stipend of \$500 (which will be issued by check and sent via postal mail) after program has concluded.

What you need to do next:

*****Please send an email to SimonsSRP@stonybrook.edu by April 5th indicating whether or not you plan to accept the offer and participate in our program.** Should you decline the offer for any reason, we may be able to extend an offer to another student who would like the opportunity to participate in our program. If you accept the offer to participate in the 2024 Simons SRP program, we will send you additional information & forms in the following weeks.

If you have any questions at all, please do not hesitate to contact us. On behalf of Stony Brook University and all the faculty who served on our selection committee, we extend our congratulations to you, and look forward to meeting and working with you this summer.

sincerely,
Karen Kernan, Director
Simons Summer Research Program

--

Simons Summer Research Program
Stony Brook Union Suite 111-06, Stony Brook University
Stony Brook, New York 11794-3257
tel: (631) 632-7114 / fax: (631) 632-4525
Email: SimonsSRP@stonybrook.edu
karen.kernan@stonybrook.edu
Simons Webpage: <http://www.stonybrook.edu/simons>

Abstract

Time Series data for medical information has the potential to revolutionize medical diagnosis by predicting diseases such as acute kidney failure months before they get diagnosed. However, current methods of doing this suffer from issues such as noisiness. This project uses the Trinetx dataset to characterize eGFR (Estimated Glomerular Filtration Rate) data, perform data imputation, test the performance of simple models as a function of the prediction range, and test time series transformer models and their performance. Data imputation was able to increase model performance by up to 3% in terms of the F1 score, and the models were able to obtain high accuracy for prediction ranges of up to 3 years. Further, transformer models were able to demonstrate effectiveness at classification tasks.

Introduction

Advances in AI have shown promise in improving early disease detection, yet challenges remain, particularly in analyzing medical time series data like eGFR and glucose screenings from EHRs. These data are often noisy, lack interpretability, and exhibit demographic-specific distributions. Our project focuses on characterizing the distribution of the data, identifying viable prediction time ranges, addressing data disparities, improving performance through imputation, and testing transformer models on medical time series data. By leveraging these models, we aim to facilitate earlier interventions and better healthcare outcomes.

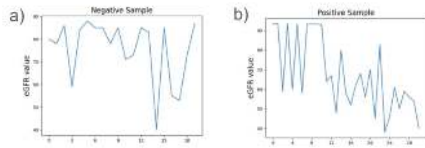


Figure 1: a) eGFR measurements for a healthy patient b) eGFR measurements up to AKF diagnosis

KEY REFERENCES

- Nie, Yuqi, et al. "A Time Series is Worth 64 Words: Long-term Forecasting with Transformers." 2023, arXiv, <https://arxiv.org/abs/2211.14730>.
- A. Y. Yildiz, E. Koç and A. Koç, "Multivariate Time Series Imputation With Transformers," in IEEE Signal Processing Letters, vol. 29, pp. 2517-2521, 2022, doi: 10.1109/LSP.2022.3224880.
- Levey, A.S., Coresh, J., Tighiouart, H. et al. Measured and estimated glomerular filtration rate: current status and future directions. *Nat Rev Nephrol* 16, 51–64 (2020). <https://doi.org/10.1038/s41581-019-0191-y>
- Hyo-Eun Kim, Hak Hee Kim, Boo-Kyung Han, Ki Hwan Kim, Kyunghwa Han, Hyeonseob Nam, Eun Hye Lee, and Eun-Kyung Kim. "Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study". en. In: *Lancet Digit. Health* 2.3 (Mar. 2020), e138–e148.

ACKNOWLEDGMENTS

I would like to thank the Simon Summer Research Program for this amazing opportunity and Dr. Yurovsky for being the Principal Investigator and guiding me through this project. Additionally, I would like to thank Dr. Tengfei Ma for providing advice throughout the project.

METHODOLOGY

1) Characterize Data Distribution

Characterized the distributions of both the negative dataset and positive dataset.

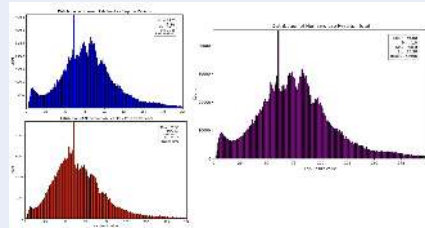


Figure 2: Lab result distributions for positive patients, negative patients, and total.

3) Simple Model Testing

Train simple models such as Random Forest, KNN, and XG Boost and test their performance.

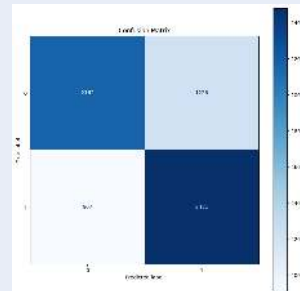


Figure 5: Confusion matrix for KNN model.

2) Data Filtering and Imputation

Data was filtered by removing patients with a low number of lab results, filtering the months such that only results before the diagnosis date are kept, and removing / imputing outliers and invalid values. Various imputation methods were used such as KNN imputation, linear interpolation, mean imputation, and spline interpolation.

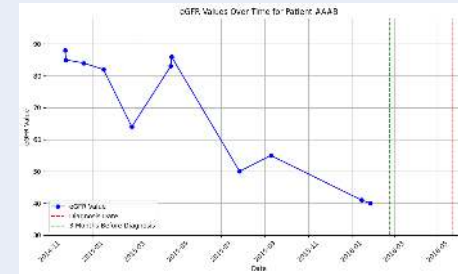


Figure 3: Lab result filtering past a certain date to simulate diagnosing 3 months ahead.

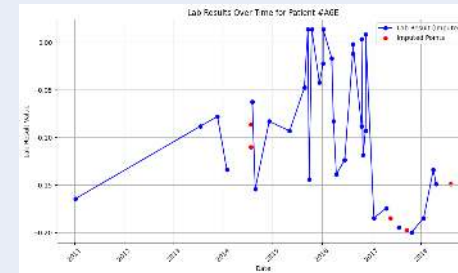


Figure 4: Linear Imputation visualized for a specific positive patient

4) Testing Transformer Architecture

- eGFR time series data is divided into fixed-length patches.
- Patches are converted into embeddings with added positional encodings to retain temporal information.
- Utilizes self-attention mechanisms to capture dependencies between different time patches.
- Multiple encoder layers process the embeddings to create a comprehensive feature representation.
- The pooled output embeddings are passed through a fully connected layer.
- The output layer classifies the time series data into the appropriate category (e.g., acute kidney failure).
- Cross-entropy loss is used for classification.
- Optimized with Adam optimizer



Figure 6: Validation and training loss plot for Patch Time Series Transformer

RESULTS

- Imputation methods such as KNN performed with 2-3% increased f1 scores.
- 5% drop in f1 score due to increased time range predictions.
- Transformers have a similar f1 score to other simple models of ~70%

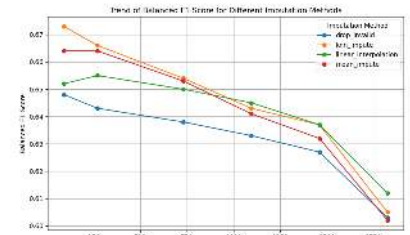


Figure 6: F1 score of random forest model as the prediction time increases. Different imputation methods are plotted.

DISCUSSION

- The relatively low drop in performance from increasing the time range suggests that the current models are not complex enough to capture all the time series information.
- The substantial improvement from imputation could lead to significant benefits for larger models, especially when applied to various datasets.
- Data imputation helps equal the distributions for negative and positive datasets and thus increases the training dataset by a large proportion.

FUTURE WORK

- Integrate various lab tests (e.g., glucose levels, blood pressure) to enhance prediction accuracy.
- Vision Time Series Transformers (ViTST) may be used to increase performance.
- Expand focus to include early detection of diseases like heart disease, diabetes, and chronic kidney disease.
- Develop demographic-specific models to improve robustness and reduce biases.
- Utilize self-supervised learning on extensive EHR data to create a comprehensive foundation model.
- Enhance interpretability by refining shapelet-based methods and validating them with clinical studies.



Thomas Li <dbbested@gmail.com>

FW: Laboratory Learning Program - Acceptance email for -Thomas Li

Lab Learning Program <lab-learning@princeton.edu>

Tue, Apr 16, 2024 at 4:15 PM

To: "dbbested@gmail.com" <dbbested@gmail.com>

Cc: "Z. Jason Ren" <zjren@princeton.edu>, "Moir J. Selinka" <mselinka@princeton.edu>, Junjie Zhu <junjiez@princeton.edu>, "Iliu@mit.edu" <liju@mit.edu>, Lab Learning Program <lab-learning@princeton.edu>

Dear Thomas,

Congratulations! We are pleased to inform you that you have been selected for the following opportunity in the Laboratory Learning Program. Participants often find that the research experience in the Laboratory Learning Program is transformative and very useful in choosing a field of study as they enter college. In this independent research opportunity, the educational benefit is closely linked to each individual's level of intellectual commitment and effort. This unique experience will allow you to be part of a world-class research endeavor.

Name of research project: ACEE-03 , Large Language Model Applications in Environmental Sustainability

Faculty mentor: Jason Ren

Faculty email: zjren@princeton.edu

Department manager: Moira Selinka

Department manager email: mselinka@princeton.edu

Department manager phone number: 609-258-8456

Lab training: Not required

To successfully enter you into the program, please complete & submit the forms on this webpage <https://scienceoutreach.princeton.edu/laboratory-learning-program-information-accepted-students-and-parentsguardians> to FACULTY, DEPT MANAGER, **SUPERVISING** STAFF, LLP emails by **April 26, 2024**.

Please connect with the researchers and department manager to arrange the participation details (for example, dates, daily hours, training).

All student participants must submit a non-confidential minimum 2-page project summary by August 20, 2024, or upon completion of the research project.

The Laboratory Learning Program **does not include** housing, transportation, social activities, or entertainment. There is no supervision provided beyond the research-related activities. Be advised that LLP students are not eligible for University housing (lease or sublease). Whether you live locally or outside of the immediate area, parents/guardians are solely responsible for safe and appropriate housing/transportation arrangements.

Faculty/department managers: if the student is not local, you need to verify they have local accommodation with a responsible adult (confirm contact information).

Sincerely,

Dr. Paryn A. Wallace

Associate Director of Science Outreach

Lynne Brown

Office Manager, Environmental Health & Safety

Laboratory Learning Program

lab-learning@princeton.edu

CC: Faculty
Department Manager
Supervising lab member
Parent

Princeton Laboratory Learning Program Report

Thomas Li, August 16, 2024

Introduction

The growing need for accurate and comprehensive data on greenhouse gas emissions from wastewater treatment facilities has never been more pressing. As the world grapples with the challenges of climate change, the role of water resource recovery facilities (WRRFs) in both contributing to and mitigating emissions has gained significant attention. However, existing protocols, such as those outlined by the IPCC, often rely on emission factors and variables that lead to inaccuracies, including the overestimation of nitrous oxide (N₂O) and underestimation of methane (CH₄) emissions. Addressing these gaps requires a comprehensive approach. We need to collect data for each wastewater treatment facility, including detailed operational parameters, process inputs, and corresponding emissions data. However, gathering the initial preexisting information for each facility can be a daunting task. This is where the web scraper, integrated with a large language model (LLM), comes into play. We want to test the hypothesis that it is possible to gather information on wastewater treatment facilities using an LLM-based web scraper. It will gather and synthesize information on wastewater treatment facilities across the United States and lay the groundwork for a more accurate inventory of emissions within this sector. There are various types of data that need to be collected for each facility. These include publicly available greenhouse gas emissions data, contact information, net zero by 2050 plan, whether the region's greenhouse gas inventory features wastewater, and whether the region's climate action plan features wastewater.

Methodology

The scraper is composed of three parts — a content finder, an interpreter, and an LLM.

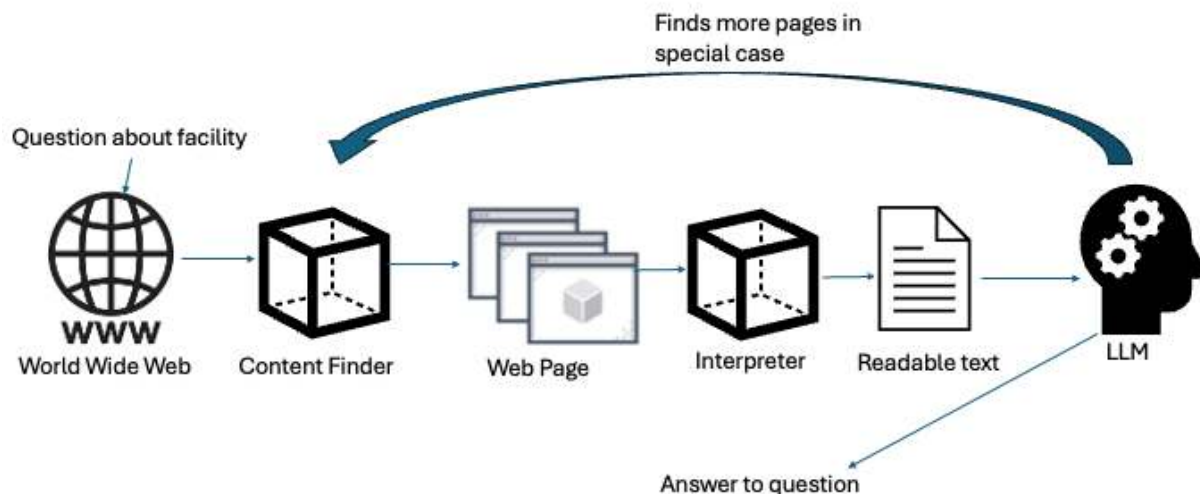


Figure 1. Displays the methodology of the scraper.

The content finder is made using the Selenium library. It searches and filters content until it finds web pages suitable for the question. Then, the web page is fed to the interpreter, which turns the web page into suitable

text that LLM can read. Different types of searches typically find other forms of web pages, so the interpreter needs to be flexible in accounting. The LLM may occasionally call back to the content finder to view a link on the web page. The system is called over every facility for specific questions.

Data

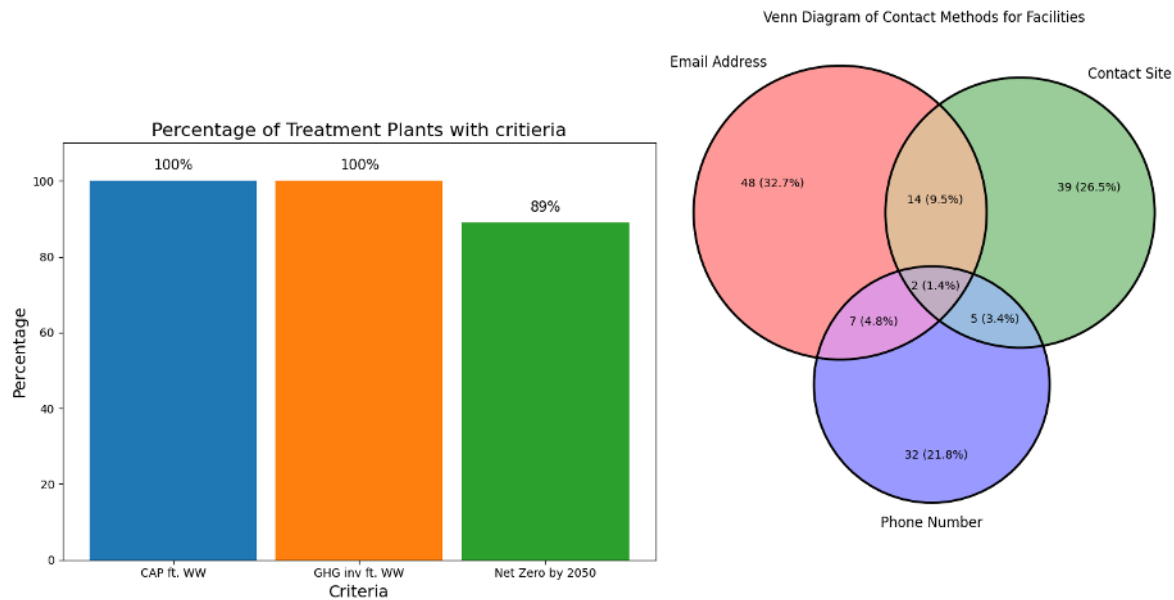


Figure 2. A) Displays the percentage of treatment plants the fulfill the given criteria. B) Displays a Venn diagram of the scraped contact information of treatment plants.

The web scraper fulfilled all tasks except finding publicly available greenhouse gas inventories.

Conclusion

We successfully created a web scraper that utilizes LLMs to find information on Wastewater Treatment Facilities. However, we need help finding data for more hidden information, such as public greenhouse gas inventories. This either shows the limitations of such methods or that such data does not exist in the public domain, and helps motivate the greater project.

Acknowledgment

I am so grateful for the support and guidance of Prof. Jason Ren and Emily Mayo. During this period, I have learned a lot, especially about the processes involved with wastewater treatment and environmental contaminants.

⚠ You must bring a printed copy of this page. We do not accept it on mobile devices.

Li, Thomas C Jan. 17, 2007



ACT Test

Sunday Feb. 07, 2021, 8:00 am

BROOKLINE HIGH SCHOOL (code 220361)

MAIN BUILDING, 1ST FLOOR ROOM 147

115 GREENOUGH ST

BROOKLINE, MA 02445

Hi Thomas! This is your ACT Test Admission Ticket. **On test day, bring acceptable photo ID***, a printout of this page, and other items listed below. To make any changes, sign in at myact.org. Thanks and good luck!

👉 SEATING ASSIGNMENTS ARE POSTED IN FRONT OF BROOKLINE HIGH SCHOOL. FIND YOUR SEATING ASSIGNMENT OUTSIDE BEFORE ENTERING THE BUILDING.

Paper Testing

Writing

Test Day Policies

- You must be at your test center by 8:00 am
- You cannot handle or activate any electronic device (including phones, smart watches and fitness watches) from test room admittance until dismissal, including breaks.
- If there's severe weather, check act.org/alerts or local radio and TV.
- You'll probably be done testing in 5-6 hours.

What to Bring

1. **Acceptable photo ID***
2. **Paper printout of this page**
3. **Pencils:** sharpened #2 lead with good erasers
4. **Permitted calculator** for use only on the math section. It is your responsibility to know whether your calculator is permitted. To find out, see act.org/calc or call 800-498-6481 for a prerecorded message.
5. **A watch** (but no smart watches or watch fitness bands) to pace yourself.
6. **Snack** to eat outside the test room on break.

What Not to Bring

- ⊘ **ANY electronic device** other than a permitted calculator.
- ⊘ Highlighters, correction fluid, colored pencils, or other aids.
- ⊘ Textbooks, foreign language or other dictionaries, scratch paper, notes or other aids or reading material.
- ⊘ Tobacco of any form.

⚠ **Test Day Match Info:** On your test day answer document, you **must** enter this data **exactly** as it appears below:

- Match Name (1st 5 letters): **LI**
- Match Number: **00601-40901**
- Date of Birth: **JAN 17 07**

* Acceptable ID Requirements

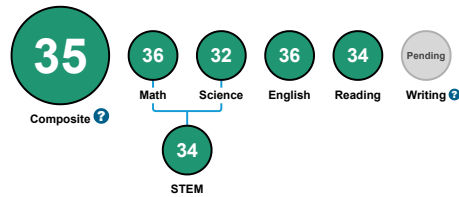
To test, you **must** bring one of these forms of ID. The name on the ID must match the name shown above on this ticket.

- Original, current official government-issued photo ID
- Current School ID in hard plastic-card format
- Otherwise, print an ACT Student ID Form at act.org/studentid
- *Talent Search students only:* Use the ACT Talent Search ID Form provided with this ticket.

Thomas's ACT Scores

Test Information Release (<https://www.act.org/publications>)

▼ February 2021



[Make Sense of Your Scores \(https://www.act.org/content/act/en/products-and-services/the-act/scores/unc\)](https://www.act.org/content/act/en/products-and-services/the-act/scores/unc)

What did Colleges and Agencies Receive?

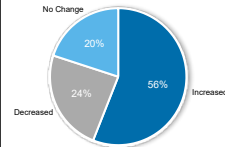
Recipients will only see the scores that you choose to send.

[MyReports](#)

What's Next?

Preparation can lead to progress! The more you study and prepare the more likely you are to improve your scores. Click on the "what can I do to improve" button for official ACT resources. You can also check out our Ideas for Progress (<https://www.act.org/content/act/en/college-and-career-readiness/benchmarks/ideas-for-progress.html>) for some activities that can lead to improvement.

Typical Composite Score of a Retest



[What can I do to improve? \(/products\)](#)

36 Math

Preparing for Higher Math (35/36)
97%

Number & Quantity (6/6)
100%

Algebra (8/8)
100%

Functions (8/8)
100%

Geometry (8/8)
100%

Statistics & Probability (5/6)
83%

Integrating Essential Skills (24/24)
100%

Modeling (16/16)
100%

32 Science

Interpretation of Data (17/19)
89%

Scientific Investigation (6/8)
75%

Evaluation of Models, Inferences & Experimental Results (11/13)
85%

36 English

Production of Writing (23/23)
100%

Knowledge of Language (11/11)
100%

Conventions of Standard English (40/41)
98%

34 Reading

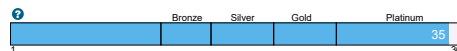
Key Ideas and Details (20/22)
91%

Craft and Structure (11/12)
92%

Integration of Knowledge and Ideas (6/6)
100%

Understanding Complex Texts ?
Below Proficient Above

Progress Toward the ACT National Career Readiness Certificate™



You are making progress towards achieving a Platinum Level on the ACT NCRC!

[Send this Score](#)

[Hide Details](#)

[How do I Compare?](#)

[Am I Ready for College?](#)

[My Interests](#)

Need Help?

See [FAQs \(/support/\)](#) or [Contact us \(https://www.act.org/content/act/en/products-and-services/the-act/the-act-test-contact-us.html\)](https://www.act.org/content/act/en/products-and-services/the-act/the-act-test-contact-us.html).

[Home](#) [Register/Buy... \(/products\)](#) [Careers, Colleges, & Majors \(/plans\)](#) [My Account \(/profile/account-information\)](#)

[Twitter](https://twitter.com/ACTStudent/) (<https://twitter.com/ACTStudent/>) [Facebook](https://www.facebook.com/theacttest/) (<https://www.facebook.com/theacttest/>) [LinkedIn](https://www.linkedin.com/company/act/) (<https://www.linkedin.com/company/act/>)
[Instagram](https://www.instagram.com/actstudent/) (<https://www.instagram.com/actstudent/>) [YouTube](https://www.youtube.com/user/TheACTTest/) (<https://www.youtube.com/user/TheACTTest/>)



Weston High School
444 Wellesley St., Weston, MA 02493
Tel: (781) 786-5800, Fax: (781)786-5829
School CEEB Code: 222355

Li, Thomas Cheng
11 Wildwood Ln, Weston, MA 02493

Page 1 of 1

Transcript

Cumulative GPA (Unweighted): 3.96

Student #: 30272
Current Grade: 11
Birthdate: 01/17/2007
Counselor: Luke Townsend

Courses Taken 2021-2022 Grade 09

Course	Final	Credit
English 9	A	5.00
Honors World History: Medieval-Early Modern	A	5.00
Honors Geometry	A	5.00
AP Computer Science	A	5.00
Advanced Projects in Java	P	1.00
Honors Physics, Lab	A	5.00
Spanish: Intermediate I	A	5.00
Concert Band	A	4.00
Digital Literacy Seminar	P	0.50
Independent Study - Developing Unity Game	P	0.50
Physical Education 9	A	2.00
Grade 9 Health A	A	0.50
Grade 9 Health B	A	0.50
June Academy	P	0.00
Total Credits:		39.00

Courses Taken 2022-2023 Grade 10

Course	Final	Credit
English 10 Honors	A-	5.00
Speech	A	1.00
AP World History	A	5.00
Honors Algebra II	A	5.00
Principles of Business	A	2.00
Honors Biology, Lab	A	5.00
Honors Chemistry, Lab	A	5.00
Spanish: Intermediate II	A	5.00
Career Exploration Seminar	A	0.50
Ind. Study - Distributing Unity Game	P	0.50
Grade 10 Health	A	1.00
PE: Lifelong Activities	A	2.00
Regional Cuisine	A	2.00
June Academy	P	
Total Credits:		39.00

Courses Taken 2023-2024 Grade 11

Course	Final	Credit
Honors American Literature	A-	5.00
AP U.S. History	A	5.00
AP Psychology	A	5.00
AP Calculus BC	A	5.00
AP Chemistry, Lab	A	5.00
AP Physics, Lab	A	5.00
Spanish: Intermediate III Honors	A-	5.00
Guidance Seminar 11	P	0.50
PE: Personal Fitness	A	2.00
June Academy	P	
June Academy	P	
Total Credits:		37.50

High School Credit Summary

Department	Credits Earned
English	16.00
History	20.00
June Academy	
Mathematics	23.00
Multidisciplinary	2.50
Other	0.00
PE/Wellness	10.00
Performing/Visual Arts	4.00
Science	25.00
World Languages	15.00
Total	115.50

Attendance

Calendar	Days Absent
22-23 Weston High School	1.0
21-22 Weston High School	8.0

Codes for courses in which a standard letter grade is not earned:

P = Pass M = Medical
I = Incomplete N = Minimum attendance requirements not met
W = Withdrawal

Please see the Curriculum Notes for descriptions of courses and levels.

UNOFFICIAL
Guidance Counselor Signature

Date Generated: 06/24/2024



Weston High School Academic Year 2024-25

444 Wellesley Street, Weston, MA 02493

Main Office (781)786-5800

Student Number: 30272

Li, Thomas

Grade: 12

2024-25 Course List

Quarter(s)	Course	Credits
Q1-Q4	AP ENGLISH	5.0
Q1-Q4	AP EUR HIS	5.0
Q1-Q4	AP STATISTICS	5.0
Q1-Q4	AP COMP SCI PRIN	5.0
Q1-Q4	AP BIO	5.0
Q1-Q4	AP SPANISH	5.0
Q3-Q4	GRAPHIC DES I	2.0
Q3-Q4	JUNIOR HEALTH	1.0
Q1-Q2	PE10-12 FITNESS	2.0

ISSUED TO:

ACADEMIC TRANSCRIPT

Thomas Li
Parchment DocumentID: TEHYWWGP

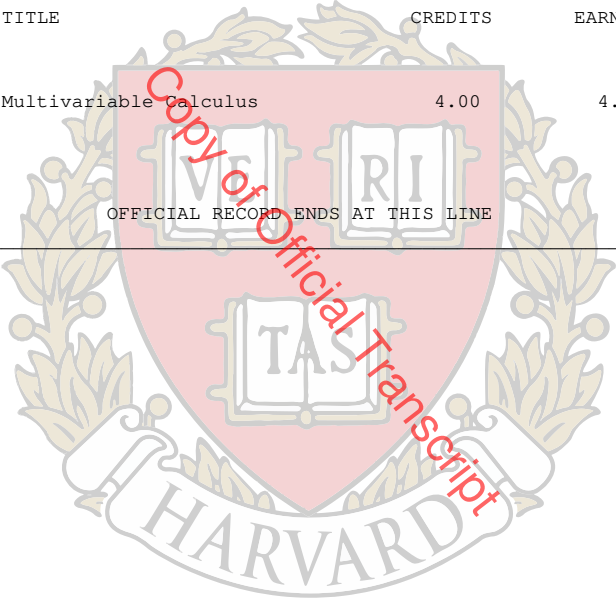
Megan E Bangs

Not valid without
official seal and signature

Name: Thomas Li
ID: @00957868

Printed: January 9, 2025
Page Number: 1

COURSE	TITLE	CREDITS	EARNED	LEVEL	GRADE	
Fall Term 2024	MATH E-21A	Multivariable Calculus	4.00	4.00	UN	A
OFFICIAL RECORD ENDS AT THIS LINE						



A CERTIFIED BLUE RIBBON VISIBLE IN ADOBE ACROBAT OR DELIVERY THROUGH PARCHMENT RECEIVE VALIDATES THIS DOCUMENT

A CERTIFIED BLUE RIBBON VISIBLE IN ADOBE ACROBAT OR DELIVERY THROUGH PARCHMENT RECEIVE VALIDATES THIS DOCUMENT

COURSE GUIDE

School code and course number



ANTH E-20 Social and Cultural Anthropology

credit value



4.00

credits earned



4.00

credit status



UN

grade



A minus

The transcript is the student's complete, official Division of Continuing Education academic record. It includes all credit and noncredit courses taken at the Harvard Extension School and the Harvard Summer School, the student's grades, all withdrawal, and certain disciplinary notations.

Degree, Diploma, and Certificate Programs: Transcripts for students admitted to these programs also include relevant degree, certificate, and diploma information, academic standing, and academic honors. Admissions to all certificate programs were discontinued in 2009 and the programs were phased out.

Citations and Professional Certificates: Beginning in 2005 students could receive a citation upon the successful completion of a proscribed set of courses. In 2011 these were renamed professional certificates. Citations and professional certificates are noted on transcripts.

Accreditation: Harvard University is fully accredited by the New England Association of Schools and Colleges. The Division of Continuing Education is accredited under Harvard University.

Privacy and Confidentiality: Division of Continuing Education policy and the Family Educational Rights and Privacy Act (FERPA) of 1974, as amended, provide students and former students certain protections and rights concerning the confidentiality of their educational records maintained by the Division of Continuing Education. This educational record is subject to FERPA, as amended, and is for official use only. It may not be released to or accessed by third parties or outside agencies without the prior written consent of the student concerned, or as allowed by law.

Harvard University School Codes: Students in degree and certificate programs may apply for "Special Student Status" and enroll in courses at other schools within the University as part of their Extension School program. Harvard employees also may apply courses taken within the University towards their Extension School program. SCHOOL CODES: BU-Business School; DN-Dental School; DS-School of Design; DV-Divinity School; E-Extension School; ED-Graduate School of Education; F-Faculty of Arts and Sciences; G-Graduate School of Arts and Sciences; GV-John F. Kennedy School of Government; LW-Law School; MD-Medical School; PH-School of Public Health; S-Summer School.

Course Title: A bracketed course title indicates the course was taken more than once and does not count toward the degree. E- and S- course numbers with the letter R indicate that the course can be repeated once for graduate credit towards the ALM degree by students who obtain approval in advance.

Credit Status: Students may enroll in courses for undergraduate (UN) credit, graduate (GR) credit, or noncredit (NC). Noncredit students are not assigned letter grades and do not receive credit for the course.

Credit Hours: One credit unit at the Division of Continuing Education is equivalent to one semester hour.

Grades: Grades reflect the quality and quantity of a student's work submitted throughout the term. Students may earn or be assigned one of the following grades and notations.

Letter grades are A, A-, B+, B, B-, C+, C, C-, D+, D, D-, and E. Non-letter grades are CR (Credit) and NC (Noncredit); PA (Pass) indicates the student passed the course (D- or better for undergraduate credit, B- or better for graduate credit); SAT (Satisfactory) indicates the course was completed satisfactorily (D- or better for undergraduate credit, B- or better for graduate credit); WA (Administrative withdrawal) is assigned to students who are administratively withdrawn in accordance with School policy.

Failing grades and notations include ABS (Absent from the final exam), E, EXD (Excluded from course), FL (Fail), INC (permanent incomplete), RQ (Required to withdraw by the Administrative Board), UNS (Unsatisfactory).

TNC indicates the ALM thesis was not completed; WD is assigned to students enrolled for undergraduate and graduate credit who withdrew from the course by the withdrawal deadline; WN is assigned to noncredit students who withdraw from the course by the deadline.

Interim grade notations include DE, EXT, INP, MU, and ***. A DE notation is assigned to students approved to take the final exam as a distance exam. EXT notation indicates the student was approved for an extension of time to complete course requirements. INP is assigned to degree students who have work in progress on their thesis, internship, or reading and research course. MU indicates the student was approved to take a make-up final exam. An asterisk *** is an interim grade notation assigned by the Registrar's Office to students with cases pending before the Administrative Board.

For details of academic policies and programs see the website or call, Academic Services at (617) 495-0977.

Student Score Report



Report Date: 10/01/2024

Thomas C. Li
121 Jericho Rd
Weston, MA 02493

AP ID/AP Number: W27V17UY

School: 222355 - Weston High School

Year taken	Name of exam	Score
2024	Calculus BC	5
	- Subscore: Calculus AB Subscore	5
2024	Chemistry	4
2024	Physics C: Electricity and Magnetism	5
2024	Physics C: Mechanics	5
2024	Psychology	4
2024	United States History	4
2023	World History: Modern	5
2022	Chinese Language and Culture	4
2022	Computer Science A	5

Awards	
2024	AP Scholar with Distinction
2023	AP Scholar



ALICE H. PEISCH
REPRESENTATIVE
14TH NORFOLK DISTRICT
WELLESLEY - WESTON - LINCOLN

The Commonwealth of Massachusetts
House of Representatives
State House, Boston 02133-1054

ASSISTANT MAJORITY LEADER

STATE HOUSE, ROOM 234
TEL. (617) 722-2783
Alice.Peisch@MAhouse.gov

February 2, 2024

Thomas Li
444 Wellesley Street
Weston, MA 02493

Dear Mr. Li:

I write to congratulate you on placing first in the Platinum and Gold Divisions, at the nation's largest cyber defense contest, the Cyberpatriot's National Youth Cyber Defense Competition. This is a significant achievement, acknowledging your cybersecurity skills and your ability to perform under pressure. You have demonstrated strong leadership communication and cooperation skills throughout this program.

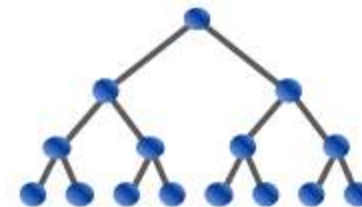
Again, congratulations, and please do not hesitate to contact me in the future if I can ever be of any assistance.

Very truly yours,

A handwritten signature in blue ink, appearing to read "AHP", with a stylized flourish at the end.

ALICE HANLON PEISCH
State Representative
14th Norfolk District

USA Computing Olympiad

[OVERVIEW](#)[TRAINING](#)[CONTESTS](#)[HISTORY](#)[STAFF](#)[RESOURCES](#)

EDIT ACCOUNT DETAILS

If you identify as female, you can be among those considered for selection as a finalist to train for the USA team attending the European Girls Olympiad in Informatics (EGOI), assuming you also list USA as your country. If this is the case, please indicate below that you are EGOI eligible.

Username:	lit	
Division:	Gold	
Email Address:	<input type="text" value="25lit@weston.org"/>	
First / Given Name:	<input type="text" value="Thomas"/>	(like John or Jane)
Last / Family Name:	<input type="text" value="Li"/>	(like Smith)
School:	<input type="text" value="Weston High School"/>	
Graduation Year:	<input type="text" value="2025"/>	(set to 9999 if you are past high school / secondary school)
Country / Region:	<input type="text" value="USA United States"/> ▼	
EGOI eligible:	<input type="text"/> ▼	
Password:	<input type="password" value="....."/>	
New Password:	<input type="password"/>	(Leave blank if you don't want to change your password)
New Password (Again):	<input type="password"/>	
<input type="button" value="Submit"/>		



Name: **Thomas C Li**
Grade: **12**
Test administration: **SAT October 5, 2024**
Tested on: **Oct 5, 2024**
Record Locator: **4098726512**

Your Scores

SAT Scores

TOTAL SCORE
1550 | 400–1600 | 99th*
Score Range: 1510-1590
3 Year Average Score (all testers): 1030

SECTION SCORES

Reading and Writing
760 | 200–800 | 99th*
Your Score Range: 730-790
3 Year Average Score (all testers): 521

Math
790 | 200–800 | 99th*
Your Score Range: 760-800
3 Year Average Score (all testers): 509

* Percentiles represent the percent of 12th grade test takers from the past 3 years who scored the same as or below you.

Score range: This is the range of scores you could possibly get if you took the SAT multiple times on different days.

Knowledge and Skills

View your performance across the 8 content domains measured on the SAT. For more information on performance score bands, visit satsuite.collegeboard.org/skills-insight.

Reading and Writing

Information and Ideas
(26% of test section, 12-14 questions)
Performance: 680-800

Craft and Structure
(28% of test section, 13-15 questions)
Performance: 680-800

Expression of Ideas
(20% of test section, 8-12 questions)
Performance: 680-800

Standard English Conventions
(26% of test section, 11-15 questions)
Performance: 680-800

Math

Algebra
(35% of test section, 13-15 questions)
Performance: 680-800

Advanced Math
(35% of test section, 13-15 questions)
Performance: 680-800

Problem-Solving and Data Analysis
(15% of test section, 5-7 questions)
Performance: 680-800

Geometry and Trigonometry
(15% of test section, 5-7 questions)
Performance: 680-800

Career Insights Snapshot

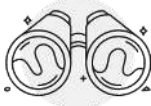
Your skills are in demand! Consider exploring these growing careers in your state that need skills like yours. **These are examples, not recommendations**, to help jump-start your career exploration.

INTEREST AREA*	CAREER EXAMPLES	MEDIAN YEARLY INCOME IN MA	NUMBER OF JOBS	JOB GROWTH IN MA**	MOST COMMON EDUCATION LEVEL
Creators	Producers and Directors	\$68,000	5,000	+6%	Bachelor's degree
Organizers	Computer and Information Systems Managers	\$174,000	28,000	+11%	Bachelor's degree
Doers	Veterinarians	\$130,000	2,400	+18%	Doctoral degree or more
Helpers	Pediatricians	\$205,000	1,100	+5%	Doctoral degree or more
Persuaders	Judges, Magistrate Judges, and Magistrates	\$208,000	1,000	+9%	Doctoral degree or more
Thinkers	General Internal Medicine Physicians	\$229,000	2,200	+7%	Doctoral degree or more

*For more information about career insights, including interest areas, visit satsuite.org/whatsnext.


**Number reflects the expected number of new job openings in 2028. Percentage reflects the expected growth in new job openings between 2023 and 2028.

Data source: Lightcast.io



Don't see a career you're interested in?

Learn more about almost 1,000 careers for free.



satsuite.org/whatsnext

© 2024 College Board.

- Led a team of 5 college undergraduate and high school students to algorithmically summarize millions of academic papers from Cornell's Arxiv.org database, using machine learning and term frequency identification.
- Developed a database and website to automatically summarize and post a bulleted description each time a new paper is published on Arxiv, leveraging recent advancements in Large Language Models and GPT.
- Organized weekly meetings to cover progress and agenda for each week, including delegating tasks for coding the database, website, and search engine

Welcome to Byte Size Arxiv

Papers made digestable

Here are some
recommened
categories:

2024-08-30

Signatures of polarized chiral spin disproportionation in rare earth nickelates

Of several proposed mechanisms, here we test the hypothesis that pairs of self-doped ligand holes spatially condense to provide local spin moments that are antiferromagnetically coupled to Ni spins. These singlet-like states provide a basis for long-range bond and spiral spin order. These spin spirals are found to preferentially couple to X-ray helicity, establishing the presence of a hitherto-unobserved macroscopic chirality. Experimentally constrained theoretical double-cluster calculations confirm the presence of an energetically stable spin-disproportionated state with Zhang-Rice singlet-like combinations of Ni and ligand moments.

[Read more →](#)

2024-08-30

Performance Analysis of Pair-wise Symbol Detection in Uplink NOMA-ISaC Systems

The results show that the derived expressions are perfectly matched with the simulation results.

[Read more →](#)

2024-08-30

Two-loop QCD corrections to Higgs radiative decay to vector quarkonium

We calculate the two-loop QCD corrections to $H \rightarrow J/\psi(\Upsilon) + \gamma$ using an optimized nonrelativistic QCD (NRQCD) approach. },
 $H \rightarrow \gamma^* \gamma$ followed by the virtual photon fragmentation into J/ψ .

[Read more →](#)

2024-08-30

Signatures of topology in generic transport measurements for Rarita-Schwinger-Weyl semimetals

Furthermore, the final expressions turn out to be insensitive to the specific values of the chemical potential and the temperature. The second scenario involves computing the magnetoelectric conductivity under the action of collinear electric (\mathbf{E}) and magnetic (\mathbf{B}) fields, representing a planar Hall set-up.

[Read more →](#)

2024-08-30

Positivity and Green's operators

In this paper we investigate the matter of just how far from being positivity preserving a positive Green's operator can be. We will also identify a broad class of Green's operators that are not necessarily

positivity preserving but have properties related to positivity preservation that one expects from positivity preserving Green's operators.

[Read more →](#)

2024-08-30

Linear constellations in primes with arithmetic restrictions

We prove analogues of the Green-Tao-Ziegler theorem on linear constellations in primes, in which the primes under consideration are restricted by certain arithmetic conditions. To illustrate this second result, we present an application in inverse Galois theory.

[Read more →](#)

2024-08-30

From free idempotent monoids to free multiplicatively idempotent rigs

A multiplicatively idempotent rig (which we abbreviate to mirig) is a rig satisfying the equation $r^2 = r$. We show that a free mirig on finitely many generators is finite and compute its size.

[Read more →](#)

2024-08-30

Quantum state testing with restricted measurements

In practice, not all measurements can be easily applied, even using unentangled measurements where each copy is measured separately. Our work bridges this gap in the literature.

[Read more →](#)

FACULTY OF SCIENCE, DEPARTMENT OF PHYSICS

CERTIFICATE OF DISTINCTION

2024 RECIPIENT

Thomas Li

WHO, IN THE ANNUAL SIR ISAAC NEWTON EXAM, PLACED 133RD IN THE
WORLD, OUT OF 3077 PARTICIPANTS AND 12TH IN UNITED STATES, OUT OF 86
PARTICIPANTS



ROGER MELKO
Exam Coordinator, Sir Isaac Newton Exam

School: *Onphysics.Com And Greater Boston
Chinese Cultural Association (Gbcca)*
Exam Date: *May 2, 2024*
Teacher: *Xiaoliang Qin*

