Geeticka Chauhan

Machine Learning for Healthcare
NLP, Computer Vision and Differential Privacy

geetickachauhan@gmail.com
people.csail.mit.edu/geeticka

EDUCATION

Massachusetts Institute of Technology — Sep 2019-June 2024
PhD in Computer Science, advisor: Peter Szolovits

Massachusetts Institute of Technology — Sep 2019
S.M. in Computer Science, advisor: Peter Szolovits

Florida International University — Jun 2017
B.S. in Computer Science, *Summa Cum Laude*

Skills: Python, C, C++, Java, Pytorch, Tensorflow, Pandas, Pax, Jax, Machine Learning

INDUSTRY INTERNSHIPS

Google Research, California — Jun 2023 - Jan 2024
Working as a Research Intern in the field of privacy-preserving machine learning. Applying differential privacy techniques to advance speech models at Google. Supervisors: Om Thakkar and Abhradeep Guha Thakurta.

Google Research, New York — May-Aug 2022
Worked as Research Intern on entity extraction of layout-rich business documents. Used techniques from sequence modeling in NLP and computer vision to expand extraction techniques for handling broken text spans and non-text entities with an impact on question answering and summarization research as well. Supervisors: Guolong Su and Vincent Perot.

Bosch Center for Artificial Intelligence (BCAI), Pittsburgh — Jun-Aug 2020
Worked as Machine Learning Research Intern, and worked on Visual Question Answering. Hosted by the team *Robust and Safe Deep Learning*. Supervisors: Heike Adel, Trung Kien Tran; collaborators: Prateek Katiyar, Leonid Boytsov

RESEARCH SUMMARY

Differentially Private Machine Learning (ongoing work)
- Researching transformer based methods for differentially private pre-training of large speech models using the DP-SGD algorithm built on top of the pax and jax frameworks
- Establishing the first baseline in the community for our project, and researching effective methods to improve privacy-utility trade-offs using novel clipping and model pruning techniques

Multimodal Machine Learning for Radiology
- Introduced a novel encoder-based transformer and ResNet based contrastive learning framework for joint training of radiology reports and chest x ray images for pulmonary edema severity prediction [11] before CLIP was introduced
- Demonstrated a 10-15% improvement in AUC over an image-only baseline using our proposed pre-training technique
- Introduced a new benchmark for the medical imaging community for predicting pulmonary edema severity [5]
- Continued exploring the multimodal direction towards using radiology report generation as a pre-training task (using sequence to sequence transformer model with ResNet), requiring only 100 labeled training examples for good performance [9]
- Did analysis comparing contrastive learning methods with radiology report generation methods to show the superiority of the latter in performance, interpretability and higher quality report generation compared to prior methods [8]

Automating relationship extraction for Biomedical data
- Introduced a CNN based model for relation extraction on the Semeval 2018 challenge, performing at rank 6 out of 28 [4]

- Highlighted the reproducibility challenge in relation extraction, along with ablation studies showing the surprising and hidden effect of pre-processing on high performance on evaluation metrics [13]. Also addressed other concerns such as dataset split bias and hyperparameter tuning strategies contributing to highest performance variations among machine learning models. Recognized by the workshop organizers as the best paper.

Robustness in machine learning for electronic health records (EHR)
- Introduced clinically aggregated features for modeling time-series electronic health records (EHR), offering robust model performance across changing hospital systems and addressing the deployability concern in machine learning for health [15, 7]. Performed experiments showing our method avoids dips as large as 30% AUC caused by applying standard time-agnostic ML methods.
- Our work has been highly well-received in the ML for health community, and is collectivelly cited over 200 times.

A new approach for entity extraction in folktales
- Introduced a novel approach using coreference resolution with support vector machines for identifying characters in folktales, which is a more specialized and challenging task in entity extraction [3, 2].

HONORS & AWARDS

MIT: Frederick (1953) and Barbara Cronin Fellowship (1 year), Graduate Woman of Excellence, Roger and Dottie Mark Award, Grace Hopper Women in Computing Scholarship, Thriving Star in EECS, CRA-W Research Travel Award, Google CS Research Mentee. GPA: 4.8/5

FIU: Presidential Scholarship (4 years), Outstanding Undergraduate in Computer Science, Banner Marshal, Dean's List (every semester), Honors College. GPA: 3.97/4

High School: Gold medalist for academic excellence for 7 consecutive years at DPS RK Puram, New Delhi, recognized as the best high school in India by Education Times. Passing out rank 13 from 1000 students with GPA 95.4/100.

PUBLICATIONS

Google scholar link: `tinyurl.com/geeticka-google-scholar`

[1] Triana Carmenate, Peeraya Inyim, Nupoor Pachekar, **Chauhan, Geeticka**, Leonardo Bobadilla, Mostafa Batouli, and Ali Mostafavi. Modeling occupant-building-appliance interaction for energy waste analysis. *Procedia Engineering*, 145:42–49, 2016.

[2] Labiba Jahan, **Chauhan, Geeticka**, and Mark Finlayson. A new approach to animacy detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1–12, 2018.

[3] Labiba Jahan, **Chauhan, Geeticka**, and Mark A Finlayson. Building on word animacy to determine coreference chain animacy in cultural narratives. In *Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference*, 2017.

[4] Di Jin, Franck Dernoncourt, Elena Sergeeva, Matthew McDermott, and **Chauhan, Geeticka**. Mit-medg at semeval-2018 task 7: Semantic relation classification via convolution neural network. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 798–804, 2018.

[5] Ruizhi Liao, **Chauhan, Geeticka**, Polina Golland, S Berkowitz, and Steven Horng. Pulmonary edema severity grades based on mimic-cxr (version 1.0. 1). *PhysioNet*, 2021.

[6] Matthew McDermott, Marzyeh Ghassemi, Nathan Hunt, Harini Suresh, **Geeticka Chauhan**, Tristan Naumann, and Peter Szolovits. Learning Physiological Decline via Random Structure Mortality Prediction Using Split RNNs. unpublished; poster at Machine Learning for Healthcare (ML4H) Workshop at NeurIPS 2017, 2017.

[7] Bret Nestor, Matthew McDermott, **Chauhan, Geeticka**, Tristan Naumann, Michael C Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Rethinking clinical prediction: Why machine learning must consider year of care and feature aggregation. In *Machine Learning for Healthcare (ML4H) Workshop 2018 at Neural Information Processing Systems (NeurIPS) Conference 2018*, 2018.

[8] Keegan Quigley, Miriam Cha, Josh Barua, **Chauhan, Geeticka**, Seth Berkowitz, Steven Horng, and Polina Golland. Bidirectional captioning for clinically accurate and interpretable models. *Submitted to AAAI*, 2023.

[9] Keegan Quigley, Miriam Cha, Ruizhi Liao, **Chauhan, Geeticka**, Steven Horng, Seth Berkowitz, and Polina Golland. Radtex: Learning efficient radiograph representations from text reports. In *MICCAI Workshop on Resource-Efficient Medical Image Analysis 2022*. ***Best paper award***. Springer International Publishing, 2022.

[10] **Chauhan, Geeticka**, Zhijing Jin, Brian Tse, Mrinmaya Sachan, and Rada Mihalcea. How good is NLP? a sober look at NLP tasks through the lens of social impact. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3099–3113, Online, August 2021. Association for Computational Linguistics.

[11] **Chauhan, Geeticka***, Ruizhi Liao*, William Wells, Jacob Andreas, Xin Wang, Seth Berkowitz, Steven Horng, Peter Szolovits, and Polina Golland. Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2020*, pages 529–539. Springer International Publishing, 2020.

[12] **Chauhan, Geeticka**, Matthew McDermott, and Peter Szolovits. A framework for relation extraction across multiple datasets in multiple domains. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 18–20, Florence, Italy, August 2019. Association for Computational Linguistics.

[13] **Chauhan, Geeticka**, Matthew B.A. McDermott, and Peter Szolovits. REflex: Flexible framework for relation extraction in multiple domains. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 30–47, Florence, Italy, August 2019. Association for Computational Linguistics.

[14] **Chauhan, Geeticka***, Elena Sergeeva*, Di Jin*, Wei-Hung Weng*, and Peter Szolovits. Explainable deep learning in healthcare: A methodological survey from an attribution view. *WIREs Mechanisms of Disease*, page e1548, 2022.

[15] Shirly Wang, Matthew BA McDermott, **Chauhan, Geeticka**, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 222–235, 2020.

TEACHING

- **Teaching Assistant**, MIT EECS — Spring 2022
  6.869: Advances in Computer Vision
- **Instructor**, MIT MISTI — Jan & Aug 2020, Jan 2022
  Machine Learning for MIT Global Startup Labs in Uruguay
- **Undergrad Teaching Assistant**, FIU — Fall 2014 & Spring 2015
  MAD 2104: Discrete Mathematics
- **Undergrad Teaching Assistant**, FIU — Spring 2014
  COP 2210: Introduction to Programming in Java

ACADEMIC ENGAGEMENTS

Invited Talks
- MIT EECS Thriving Stars Annual Research Summit — May 2023
- Clinical Informatics Lecture series, — April 2020
  Harvard Medical School Biomedical Informatics division

Conference Oral Presentations
- REflex: Flexible Framework for Relation Extraction in multiple domains [13], — Jul 2019
  BioNLP Workshop at ACL 2019

- Joint Modeling of Chest X Ray images and radiology reports for pulmonary edema severity prediction [11], MICCAI 2020   Oct 2020

Conference Poster Presentations
- MICCAI 2020 [11], Widening NLP 2019 at ACL 2019 [13], CRA-W Grad Cohort 2019 [13], Women in Data Science Cambridge 2019 [13], MIT Schwarzman College of Computing Launch 2019 [13], Semeval 2018 at NAACL-HLT 2018 [4], Widening NLP 2018 at NAACL-HLT 2018 ([3] presented again)

Conference Workshop Organizing
- Women in Machine Learning Workshop at NeurIPS 2021
- NLP for open source software at EMNLP 2020 and 2023
- ML across MIT workshop in Feb 2020

Others
- Travel awards and volunteering: Volunteer at NeurIPS 2019, ICLR 2020, won travel grants to attend MICCAI 2020 (from MICCAI National Institute of Health Awards), ACL 2019 (from Widening NLP 2019 workshop), NAACL 2019 (from Widening NLP 2018 workshop)
- Reviewing: ACL 2023, ICLR 2023, NeurIPS 2020-21, EMNLP 2020, ICML 2020-23 (identified as top 33% reviewer in 20), MLHC 2019-21, NAACL-HLT 2019+21, EACL 2020, AMIA 2019-20 summit, ML4H 2018-19 at NeurIPS and WiML 2018 workshop at NeurIPS
- Served as the MIT EECS booth student liaison for student recruitment at the ACM Tapia Conference for improving diversity in computing, as well as the Society for Women Engineers workshop in 2022

NEWS
- MIT News: Learning the ropes and throwing lifelines
- MIT News: Anticipating heart failure with machine learning
- MIT CSAIL News: Anticipating heart failure with machine learning with press mentions at Venture-Beat, Health IT Analytics, New Atlas, Aunt Minnie, Health Imaging, Medical Xpress and Neptune AI
- MIT News: Dreaming big in a small country
- FIU News: Higher Education gets a reboot
- MIT News: Serving up brunch for the graduate community
- Behind the scenes of a contact tracing study
- The Hindu (Indian National Newspaper): Destination Florida

SERVICE
- ML across MIT Funding Chair   2019-20
- MIT Graduate Ring Committee   2018-2019
- MIT Presidential Advisory Council   2019-2020
- President, MIT Sidney Pacific Grad Community   2019-2020
- Vice President, MIT Sangam   2018-2019
- Instructor, MIT Online Science, Technology and Engineering Community   Aug 2019
- Instructor, Google IgniteCS at FIU   2016-2017
- Instructor, DPS Shiksha Kendra in New Delhi   2012-2013