Samir Gupta
Research Instructor
Georgetown University

Email : sg1446@georgetown.edu
Phone : +1-302-602-6057
NIH Biosketch

Research Interests

Natural Language Processing (NLP), Machine Learning (ML) and its application to the Biomedical and Clinical Domain, Clinical NLP, Clinical Informatics, Health Data Science, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG)

Work Experience

- **Research Instructor/Faculty** — April 2021 – present
  Innovation Center for Biomedical Informatics & Dept. of Oncology, Georgetown University — *Washington, DC*
- **Post-Doctoral Fellow, Informatics** — April 2019 – April 2021
  Innovation Center for Biomedical Informatics, Georgetown University — *Washington, DC*
- **Research and Teaching Assistant** — Sep. 2011 – March 2019
  Dept. of Computer and Information Sciences, University of Delaware — *Newark, DE*
- **Software Engineer** — Oct. 2009 - July 2011
  Interra Systems — *Kolkata, India*

Education

- **Ph.D. in Computer and Information Science** — Nov 2020
  University of Delaware; GPA: 4.0 — *Newark, DE*
  - **Dissertation Topic**: Extraction of knowledge for microRNAs and genes: Extracting connections through Association, Involvement, and Regulation
- **M.S. in Computer and Information Science** — May 2014
  University of Delaware; GPA: 4.0 — *Newark, DE*
- **B.E. (honors) in Computer Science and Engineering** — July 2009
  Jadavpur University; GPA: 8.78/10.0 — *Kolkata, India*

Research Experience

- **Natural Language Processing for Clinical Text**: Development of NLP tools to extract various clinical phenotypes and entities, such as outcomes, adverse events, SDOH, and medication information from clinical notes.
  - Developing an biomedical Named Entity Recognition (NER) pipeline based on **ClinicalBERT** to identify patients with cancer-treated Immune Checkpoint Inhibitors (ICIs) as having developed **adverse events** based on their longitudinal set of clinical notes. We used the "Georgetown-Lombardi Immuno-Oncology Registry", which contains clinical information along with treatment outcomes for about 1000 cancer patients from MedStar Health network hospitals to inform the ML models.
  - Collaborated with scientists at the Frederick Cancer National Labs to develop a tool to extract **NSAID medication usage** information such as drug type, frequency, route, and strength from clinical notes, which is being used in an observational study to link NSAID use with overall survival of lung cancer patients.
- **Assisting Biocuration through Automation**: Development of ML-driven NLP tools to extract literature evidence to assist variant pathogenicity curation.
  - Developed a text-mining tool (MACE2k) to extract associations between variants, diseases, and drug responses from full-length articles to assist in variant pathogenicity interpretation activities of ClinGen and CIViC. Investigating state-of-the-art ML techniques such as **distant supervision** and **transfer learning** to learn from a small curated data set and a large "noisy" automatically labeled data.
  - Developing eMIND, a Relation Extraction (RE) system that supports the automatic extraction of impacts of variants on protein proprieties in Alzheimer's disease from literature. We leverage langauge models (BioBERT, **PubMedBERT**) and investigate the utility of **Retrieval-Augmented Generation (RAG)** to improve RE system by providing the relations identified by BERT-based models to Large Language Models (LLMs) as enhanced context.

- **Machine Learning for Observational Studies**: Developing ML predictive models to study correlations between clinical features and health outcomes
  - Developed predictive models to study **response to immunotherapy** in NSCLC patients based on demographics, lab results, medications, and molecular test results.
  - Involved in building predictive models to detect **suicide ideation** among US veterans based on their audio recordings collected through a mobile app. This research project (collaboration with the VA and Department of Psychiatry) entailed the development of an NLP and ML pipeline to screen for suicidality based on acoustic features of speech and linguistic features of the transcribed audios.
  - Worked on a project to detect patients who underwent upper extremity surgeries that are at substantial risk of developing **opioid use disorder**. This study is in collaboration with the Curtis National Hand Center, Baltimore, and involves more than 2000 patients who underwent surgery at the MedStar Health System.

Teaching Experience

- **Georgetown University** Washington, DC
  Instructor/Teaching Faculty, Health Informatics and Data Science (HIDS) *Sep. 2019 - present*
  - Assisting in curriculum design, instruction, and assessment of different courses of the HIDS Graduate program.
    Instructor for HIDS 6001: Massive Health Data Fundamentals
    Instructor for HIDS 7006: AI/Machine Learning for Health Applications
    Co-mentor: HIDS Capstone Project
- **University of Delaware** Newark, DE
  Teaching Assistant, Computer Science *Sep. 2011 - May 2013*
  - The nature of work involved grading assignments and holding office hours. TA for Data Structures, Introduction to Algorithms, Introduction to Computer Science, Logic, and Machine Learning

Awards and Professional Activities

- **Best Research Paper Award**: International Conference on Program Comprehension (ICPC), 2013
- **Best Research Paper Award**: International Conference on Mining Software Repositories (MSR), 2013
- **Professional Development Awards**: Received travel awards to attend ICPC 2013, SMBM 2014, BioNLP 2015, 2017
- **Program Committee Member**: Mid-Atlantic Student Colloquium on Speech, Language and Learning 2018, Student Paper Competition AMIA 2021, 2022, 2023, 2024 Annual Symposium
- **Workshop Organizing Member**: OncoMX Community Workshop 2018
- **Student Member**: American Medical Informatics Association (AMIA)
- **Professional Membership**: International Association for the Study of Lung Cancer (IASLC)
- **Journal Reviewer**: PLOS One, Bioinformatics Advances, The Journal of Biological Databases and Curation
- **Conference Reviewer**: AMIA 2020 Annual Symposium, AMIA 2022 Informatics Summit

Technical Skills

- **Languages**: Python, R, C/C++, Java, Ruby, Perl, Lisp, Shell Scripting
- **Web Development**: HTML, CSS, Javascript, Django, Flask
- **Tools**: Docker, Protocol Buffer, MongoDB, scikit-learn, Tensor Flow, Keras, Hugging Face, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG)

Publications

Yili Zhang, Jia Li Dong, Bai Xue, Yanbao Xiong, **Samir Gupta**, Maarten Van Segbroeck, Nawar Shara, Peter McGarvey. Exploring the Utilization of Synthetic Data in Unsupervised Clustering for Opioid Misuse Analysis. *American Medical Informatics Association (AMIA) Annual Symposium*, Nov 2024, Pages 1-12

[Preprint] **Samir Gupta**, Xihan Qin, Qinghua Wang, Julie Cowart, Hongzhan Huang, Cathy H Wu, K Vijay-Shanker, Cecilia N Arighi. eMIND: Enabling automatic collection of protein variation impacts in Alzheimer's disease from the literature. *bioRxiv* 2023.09.07.556602; doi: https://doi.org/10.1101/2023.09.07.556602

Zhang Y, Ari SL, **Gupta S**, et al. Enhancing patient safety in melanoma treatment: harnessing machine learning for predicting immune-related adverse events. *Journal for ImmunoTherapy of Cancer* 2023;11:doi: 10.1136/jitc-2023-SITC2023.1317

Roszik J, Lee JJ, Wu YH, Liu X, **Gupta S** et al., Dmitrovsky E. Real-World Studies Link Nonsteroidal Anti-inflammatory Drug Use to Improved Overall Lung Cancer Survival. *Cancer Res Commun.* 2022 Jul;2(7):590-601. doi: 10.1158/2767-9764.crc-22-0179. Epub 2022 Jul 6. PMID: 35832288; PMCID: PMC9273107.

Gupta S, Belouali A, Shah NJ, Atkins MB, Madhavan S. Automated Identification of Patients With Immune-Related Adverse Events From Clinical Notes Using Word Embedding and Machine Learning. *JCO Clinical Cancer Informatics* 2021 May;5:541-549.

Belouali A, **Gupta S**, Sourirajan V, Yu J, Allen N, Alaoui A, Dutton MA, Reinhard MJ. Acoustic and language analysis of speech for suicidal ideation among US veterans. *BioData Mining* 2021 Feb 2;14(1):11.

Roychowdhury D, **Gupta S**, Qin X, Arighi CN, Vijay-Shanker K. emiRIT: a text-mining-based resource for microRNA information. *Database* (Oxford). 2021 May 28;2021:baab031.

S. Gupta, A. Belouali, N. Shah, M. Atkins, S. Madhavan. Automated Identification of Patients with Immune-related Adverse Events from Clinical Notes using Machine Learning. *AMIA 2020 Annual Symposium*

Rao S, Pitel B, Wagner AH, Boca SM, McCoy M, King I, **Gupta S** et al. Collaborative, Multidisciplinary Evaluation of Cancer Variants Through Virtual Molecular Tumor Boards Informs Local Clinical Practices. *JCO Clinical Cancer Informatics* 2020;4: 602–613.

Dingerdissen HM, Bastian F, Vijay-Shanker K, Robinson-Rechavi M, Bell A, Gogate N, **Gupta S**, et al. OncoMX: A Knowledgebase for Exploring Cancer Biomarkers in the Context of Related Cancer and Healthy Data. *JCO Clinical Cancer Informatics* 2020 Mar;4:210-220.

Hu Y, Dingerdissen H, **Gupta S**, Kahsay R, Shanker V, Wan Q, et al. Identification of key differentially expressed MicroRNAs in cancer patients through pan-cancer analysis. *Computers in Biology and Medicine.* 2018;103: 183–197

Gupta S, Dingerdissen H, Ross KE, Hu Y, Wu CH, Mazumder R, et al. DEXTER: Disease-Expression Relation Extraction from Text. *Database* (Oxford). 2018 Jan 1;2018:bay045.

Gupta S, Mahmood ASMA, Ross K, Wu C, Vijay-Shanker K. Identifying Comparative Structures in Biomedical Text. *BioNLP* 2017. Stroudsburg, PA, USA: Association for Computational Linguistics; 2017.

Gupta S, Ross KE, Tudor CO, Wu CH, Schmidt CJ, Vijay-Shanker K. miRiaD: A text mining tool for detecting associations of micrornas with diseases. *Journal of Biomedical Semantics.* 2016;7: 9

Peng Y, **Gupta S**, Wu C, Shanker V. An extended dependency graph for relation extraction in biomedical texts. Proceedings of *BioNLP* 15. 2015; 21–30.

Gupta S, Tudor CO, Wu CH, Schmidt CJ, Vijay-Shanker K. Automatically Identifying Biological Functions of microRNAs from the Literature. 6th International *Symposium on Semantic Mining in Biomedicine* (SMBM 2014). 2014. pp. 75–78.

Howard MJ, **Gupta S**, Pollock L, Vijay-Shanker K. Automatically mining software-based, semantically-similar words from comment-code mappings. Proceedings of the 10th Working *Conference on Mining Software Repositories.* IEEE Press; 2013. pp. 377–386.

Gupta S, Malik S, Pollock L, Vijay-Shanker K. Part-of-speech tagging of program identifiers for improved text-based software engineering tools. 21st *International Conference on Program Comprehension* (ICPC), IEEE; 2013. pp. 3–12.

[Preprint] **Gupta S**, Rao S, Miglani T, Iyer Y, Lin J, Saiyed AM, et al. MACE2K: A Text-Mining Tool to Extract Literature-based Evidence for Variant Interpretation using Machine Learning. *bioRxiv.* 2020. p. 2020.12.03.409094

References available on request.