

Sonish Sivarajkumar

Email: sonish.sivarajkumar@pitt.edu Phone: +1 412 478 8959

[Website](#) | [LinkedIn](#) | [GitHub](#) | [Google Scholar](#)

RESEARCH INTERESTS

Natural Language Processing, Information Retrieval, Information Extraction, Few/Zero-shot Learning, Representation Learning, Clinical Large Language Models(cLLMs), Biomedical Informatics, Electronic Health Records(EHR), Real World Evidence(RWE)

EDUCATION

PhD in Intelligent Systems

University of Pittsburgh | 2021-Present

Pittsburgh, PA

Major: Informatics

Master's in Intelligent Systems

University of Pittsburgh | 2021-2022

Pittsburgh, PA

Major: Informatics

Bachelor's in Electrical Engineering

APJ Abdul Kalam Technological University | 2016-20

India

Government Engineering College - Thrissur

EXPERIENCE

Graduate Student Researcher | August 2021 – Present

University of Pittsburgh, Intelligent Systems Program

- Areas: AI in Medicine, Information Retrieval, Information Extraction, few-shot/zero-shot learning, clinical NLP, Large Language Models, Patient Representation Learning
- Working on foundational models and language models for generative patient representation using soft prompting and zero-shot learning on generative patient models. Generative patient models are models that can create new patient data or embeddings based on some input data or conditions
- Applying NLP techniques based on regular expressions, machine learning, and deep learning to extract relevant information from large-scale Electronic Health Records (EHR) and clinical notes for clinical knowledge discovery
- Investigating novel clinical embeddings using deep neural networks and Large Language Models (LLM) for clinical outcome prediction
- Exploring Zero/Few-shot learning methods such as prompt learning, Siamese Neural Networks (SNN) for Healthcare NLP applications to shift the paradigm from “Deep learning” to “Deep thinking”
- Conducting end-to-end disease studies from EHRs using clinical NLP
- Implementing clinical machine learning systems using K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), Logistic Regression, Recurrent neural network (RNN), LSTM, Generative adversarial networks (GAN), AutoEncoders (AE), and fine-tuned Deep learning PLMs (Pre-trained Language Models) like BERT, BioBERT, ClinicalBERT, etc.
- Collaborating with oncologists to develop NLP and ML algorithms to predict immunotherapy response and metastases prediction on lung adenocarcinoma patients
- Worked on building AI tools for Cancer Genomics and Spatial Transcriptomics on Visium breast cancer data.

Data Science Research Intern | May – August 2023

Merck & Co., Inc. , Philadelphia

- Performed data analysis using EHR and claims data to assess the compliance of oncology providers to NCC guidelines and the quality of care for patients with early-stage breast cancer
- Calculated the BRCA testing rates and determined the timing and location of BRCA testing for patients eligible for Lynparza, a targeted therapy that requires positive BRCA testing
- Applied NLP techniques to extract biomarker information from structured and unstructured data sources using Merck internal datasets, claims data, IQVIA and Syapse datasets
- Experimented with various language models such as BERT, BioBERT, Clinical BERT, etc. to process and analyze unstructured data such as clinical notes and reports
- Investigated the use of Google Trends to analyze the vaccination trends and the public interest in BRCA testing

gRED AI Predictive Analytics Research Intern | May – July 2022

Genentech (Roche), San Francisco

- Part of the Early Clinical Development Informatics(ECDi) team working on Clinical Operations in trial design, building predictive tools, and improving the drug and target/biomarker discoveries.
- Developed predictive clinical trials site recommendation tool, using advanced AI and NLP techniques.
- Responsible for developing a vector space model for Roche internal clinical trials sites and PIs (datasets: Citeline, AACT, ClinicalTrials.gov, Roche Internal data)
- Implemented and tested this clinical trial site embedding-based Information Retrieval system, with primary focus on Diversity and Inclusion.

Data Science Intern | August 2019 – May 2020

Fractal, India

- Areas: Risk Analytics, Customer Analytics, Data Analytics, Deep Learning
- Worked on Credit risk modeling and risk analytics for one of the largest financial companies in India.

Co-Founder and Head of Data Science | August 2017 – May 2019

StartUp India

- Areas: Intelligent Systems, Machine Learning, Deep Learning, Robotics
- Head of the AIML team of 4 members. Developed a set of AI and Robotics solutions on Healthcare, which was acquired by Rajagiri Medical College, Kochi, India.

PUBLICATIONS

- **Sivarajkumar S**, Mohammad HA, Oniani D, Roberts K, Hersh W, Liu H, He D, Visweswaran S, Wang Y. "Clinical Information Retrieval: A literature review" Under Review(2023)
- **Sivarajkumar S**, Wang Y. "Evaluation of Healthprompt for Zero-shot Clinical Text Classification" IEEE International Conference on Healthcare Informatics(2023)
- **Sivarajkumar S**, Wang Y. "A Counterfactual-based Explanation Framework for Large Language Models in Clinical Natural Language Processing." AMIA Annual Symposium(2023)
- **Sivarajkumar**, **Sonish**, and Yanshan Wang. "HealthPrompt: A Zero-shot Learning Paradigm for Clinical Natural Language Processing." *arXiv preprint arXiv:2203.05061* (2022). AMIA Annual Symposium 2022, as one of 8 finalists in Best Paper competition 2022.
- **Sivarajkumar, S.**, Viggiano, S., Oniani, D., Visweswaran, S., & Wang, Y. (2022). Extraction of

Sleep Information from Clinical Notes of Alzheimer's Disease Patients Using Natural Language Processing.

- Oniani, David, **Sonish Sivarajkumar**, and Yanshan Wang. "Few-Shot Learning for Clinical Natural Language Processing Using Siamese Neural Networks." *JMIR AI* (2023).
- Koyilot, Mufeeda C., Priyadarshini Natarajan, Clayton R. Hunt, **Sonish Sivarajkumar**, Romy Roy, Shreeram Joglekar, Shruti Pandita et al. "Breakthroughs and Applications of Organ-on-a-Chip Technology." *Cells* 11, no. 11 (2022): 1828.
- April Sagan, **Sonish Sivarajkumar**, Hatice Osmanbeyoglu "Computational methods for delineating spatially informed cell context-specific regulatory programs." *UPMC Cancer Retreat* 2021.

CONFERENCE PRESENTATIONS

- **Sivarajkumar, Sonish**, and Yanshan Wang. "HealthPrompt: A Zero-shot Learning Paradigm for Clinical Natural Language Processing. *Paper Presentation, AMIA Symposium* 2022.
- April Sagan, **Sonish Sivarajkumar**, Hatice Osmanbeyoglu "Computational methods for delineating spatially informed cell context-specific regulatory programs." *UPMC Cancer Retreat* 2021

TALKS

- Clinical NLP and Few/Zero-shot learning for Clinical Text Extraction. *Presented at: University of California-San Francisco (UCSF) Seminar series, August 2022*
- Few-shot and zero-shot Learning for Clinical Information Extraction. *Presented at: Merck Text Mining Task Force Seminar series, August 2022*
- Guest lecture on 'Programming in R' in Foundations of Health Informatics, University of Pittsburgh; *June 2022.*
- Explainable Natural Language Processing(NLP). *Presented at: Department of Biomedical Informatics, University of Pittsburgh; April, 2022.*
- Zero-Shot Learning for Clinical Natural Language Processing. *Presented at: Intelligent Systems Program AI Forum, University of Pittsburgh; February, 2022.*
- Guest lecture on 'Natural Language Processing' in Foundations of Health Informatics, University of Pittsburgh; *February 2022.*

SKILLS AND INTERESTS

Skills and Interests: Machine Learning, Deep Learning, Natural Language Processing, Information Retrieval, Information Extraction, ETL of Electronic Health Records data, Large Language Models, Explainable AI, Literature-based Discovery(LBD), Representation Learning

Languages: Python, R, Java, C, SQL, Git

Technologies: Docker, AWS, Joyent Triton Cloud, CI/CD, Data Engineering(ETL-Spark-Hive)

Libraries: Transformers, NLTK, Spacy, Pandas, Scikit-learn, Jupyter, Keras, Networkx, Tensorflow, Pytorch, Stellargraph, OpenPrompt, Pyspark

OTHER PROFESSIONAL ACTIVITIES

Editorial Activities

Journal of the American Medical Informatics Association (JAMIA) Student Editorial Board
Member | 2022-2023

Peer Review

ACL 2023 (Association of Computational Linguistics)| 2023
06/09/22 Sonish Sivarajkumar, University of Pittsburgh

ICHI 2023 (IEEE International Conference on Healthcare Informatics) | 2023
Journal of Healthcare Informatics Research(JHIR) | 2023
ICHI 2022 (IEEE International Conference on Healthcare Informatics) | 2022
ICML 2022 (International Conference on Machine Learning) | 2022
LREC 2022 (Conference on Language Resources and Evaluation) | 2022

Workshops

Publication chair EBAIC 2023 (International Workshop on Ethics and Bias of Artificial Intelligence in Clinical Applications) | *Houston, 2023*
Co- organizing the AMIA 2022 NLP Working Group Pre-Symposium | *Washington,DC, 2022*

Volunteering

Translational Bioinformatics Year-in-Review team, AMIA Informatics Summit | 2021,2022,2023
Student Volunteer, AMIA Annual Symposium 2022

Memberships

American Medical Informatics Association (AMIA) | 2020-Present
International Society of Computational Biology (ISCB) | 2021-Present
Institute of Electrical and Electronics Engineers (IEEE) | 2019-Present

AWARDS

- Fellowship – School of Computing and Information, University of Pittsburgh | 2021-2022
- Fellowship – School of Computing and Information, University of Pittsburgh | 2022-2023
- AMIA 2022 Best Paper Award Finalist | 2022