Sonish Sivarajkumar

Email: sonish.sivarajkumar@pitt.edu **Phone:** +1 412 478 8959
Website | LinkedIn | GitHub | Google Scholar

RESEARCH INTERESTS

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

EDUCATION

PhD in Intelligent Systems
Thesis: Clinical Natural Language Processing(NLP) and Predictive Modelling using Large Language Models(LLMs)
University of Pittsburgh | 2021-2025 Pittsburgh, PA
Major: Intelligent Systems Program – Informatics Track

Master's in Intelligent Systems
University of Pittsburgh | 2021-2022 Pittsburgh, PA
Major: Intelligent Systems Program – Informatics Track

Bachelor's in Electrical Engineering
APJ Abdul Kalam Technological University India
(Government Engineering College – Thrissur)| 2016-2020

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, Representation Learning, SNOMED, OHDSI tools, OMOP data models, clinical data standards such as ICD, RxNorm, SNOMED, and LOINC and other biomedical ontologies
Languages: Python, R, Java, C, SQL, Git
Technologies: Docker, AWS, Joyent Triton Cloud Computing, CI/CD, Data Engineering (ETL-Spark-Hive)
Libraries: Transformers, NLTK, Spacy, Pandas, Scikit-learn, Jupyter, Keras, Networkx, Tensorflow, Pytorch, Stellargraph, OpenPrompt, Pyspark, Hadoop, Langchain
Datasets: Real World Data- IQVIA, Mckesson, Citeline, Syapse, AACT, and LexisNexis
EHRs: MIMIC, UPMC Hospital Cancer Registry, Cerner, EPIC, ARIA

EXPERIENCE

NLP Research Scientist | ***May 2024 – present***
Eli Lilly & Co, Indianapolis

- Designed, developed and evaluated a generative AI-based system using GPT-4, LLAMA, and Mixtral to enhance clinical decision-making.
- Built NLP pipelines with LLMs for accurate extraction of cardiovascular events from EHRs and clinical trial data.
- Implemented multi-agent systems to automate and improve the accuracy of cardiovascular outcome adjudication in clinical trials.

Research Assistant | *August 2021 – June 2025*
University of Pittsburgh, Intelligent Systems Program
UPMC Hillman Cancer Center, Pittsburgh

- *Areas: AI in Medicine, Information Retrieval, Information Extraction, few-shot/zero-shot learning, clinical NLP, Large Language Models, Foundational Models, Patient Representation Learning, Generative AI*
- Working on building advanced foundational models and LLMs for clinical domain, using high-performance GPU clusters.
- Collaborating with oncologists to develop NLP and ML algorithms to predict immunotherapy response and metastases prediction on lung adenocarcinoma patients.
- Applying NLP techniques based on regular expressions, machine learning, and deep learning to extract relevant information from large-scale EHR and clinical notes for clinical knowledge discovery and decision making.
- Investigating novel clinical embeddings using deep neural networks and Large Language Models (LLM) for clinical outcome prediction. Methods include SFT (Supervised Fine-Tuning), RLAIF (Reinforcement Learning with AI Feedback), prompt engineering, Retrieval Augmented Generation (RAG), and data synthesis techniques.
- Exploring Zero/Few-shot learning methods such as prompt learning, Siamese Neural Networks (SNN) for Healthcare NLP applications.
- Conducting end-to-end disease studies from EHRs using clinical NLP
- Building clinical machine learning systems using K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), Logistic Regression, Recurrent neural network (RNN), Generative Adversarial Networks (GAN), AutoEncoders (AE), and fine-tuned LLMs like LLAMA(LoRa Instruction tuning), GPT, Mistral, BERT, BioBERT, ClinicalBERT,etc.

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, ClincalTrials.gov, Roche Internal data)

- Implemented and tested this clinical trial site embedding-based Information Retrieval system, with primary focus on Diversity and Inclusion.

Data Scientist-II | ***May 2020 – August 2021***
IQVIA, IQVIA AI Center of Excellence(CoE), India

- *Areas: Machine Learning, Big Data, Time Series Analysis, Health Care Analytics, NLP*
- Worked on "Country Patient Analytics" project, which is a big data tool for doing custom analytics on clinical and patient level RWE and EHR data.
- Worked on building Clinical AI and analytics systems using Real World Evidence(RWE) and Electronic Health Records(EHR) data
- Led a team of 4 for completing an end-to-end Clinical trials pipeline automation project using NLP and deployed the application in IQVIA's private cloud.
- Worked on a POC for segmentation and targeting of Healthcare providers (HCPs)
- Led Apache airflow migration of the big data scheduler and cloud integration and deployment of an AI and Analytics platform

PUBLICATIONS

- Wu, Xizhi, David Oniani, Zejia Shao, Paul Arciero, **Sonish Sivarajkumar**, Jordan Hilsman, Alex E. Mohr et al. "A Scoping Review of Artificial Intelligence for Precision Nutrition." *Advances in Nutrition (2025)*
- **Sivarajkumar, Sonish**, Subhash Edupuganti, Manisha Bhattacharya, David Lazris, Michael Davis, Yufei Huang, and Yanshan Wang. "Automating the detection of treatment progression in patients with lung cancer using large language models." *Journal of Clinical Oncology (2024)*
- Ji, Yuelyu, Wenhe Ma, **Sonish Sivarajkumar,** Hang Zhang, Eugene Mathew Sadhu, Zhuochun Li, Xizhi Wu, Shyam Visweswaran, and Yanshan Wang. "Mitigating the risk of health inequity exacerbated by large language models." *NPJ Digital Medicine (2025)*.
- **Sivarajkumar S**, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. "An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing", *Journal of Medical Internet Research-Medical Informatics (JMIR-MI 2024)*
- **Sivarajkumar S**, Gao F, Denny PE, Aldhahwani BM, Bove A, Visweswaran S, Wang Y. "Extracting Physical Rehabilitation Exercise Information from Clinical Notes: Algorithm Development and Validation Study", *Journal of Medical Internet Research-Medical Informatics (JMIR-MI 2024*)
- **Sivarajkumar, S.,** Viggiano, S., Oniani, D., Visweswaran, S., & Wang, Y. "Extraction of Sleep Information from Clinical Notes of Alzheimer's Disease Patients Using Natural Language Processing", *Journal of Americal Medical Informatics Association (JAMIA 2024)*
- **Sivarajkumar, S.,** Edupuganti, S., Bhattacharya, M., Lazris, D., Davis, M., Huang, Y., & Wang, Y., "Automating the detection of treatment progression in patients with lung cancer using large language models", *Journal of Clinical Oncology (JCO 2024)*
- Ji, Y., Li, Z., Meng, R., **Sivarajkumar, S**., Wang, Y., Yu, Z., Ji, H., Han, Y., Zeng, H. and He, D, "RAG-RLRC-LaySum at BioLaySumm: Integrating Retrieval-Augmented Generation and Readability Control for Layman Summarization of Biomedical Texts.", *IEEE International Conference on Healthcare Informatics (ICHI 2024)*
- Tam, T. Y. C., **Sivarajkumar, S**., Kapoor, S., Stolyar, A. V., Polanska, K., McCarthy, K. R., ... & Wang, Y.,. "A Literature Review and Framework for Human Evaluation of Generative Large Language Models in Healthcare.", *NPJ Digital Medicine(2024)*
- Gao, F., Zhang, X., **Sivarajkumar, S**., Denny, P., Aldhahwani, B., Visweswaran, S., ... & Wang,

Y. , "Precision Rehabilitation for Patients Post-Stroke based on Electronic Health Records and Machine Learning." , *Under Review*.

- **Sivarajkumar S**, Mohammad HA, Oniani D, Roberts K, Hersh W, Liu H, He D, Visweswaran S, Wang Y. "Clinical Information Retrieval: A literature review",
  Journal of Healthcare Informatics Research (JHIR 2023)
- **Sivarajkumar S,** Huang Y, Wang Y., "Fair Patient Model: Mitigating Bias in the Patient Representation Learned from the Electronic Health Records",
  Journal of Biomedical Informatics (JBI 2023)
- **Sivarajkumar S**, Wang Y. "Evaluation of Healthprompt for Zero-shot Clinical Text Classification", *IEEE International Conference on Healthcare Informatics (ICHI 2023)*
- **Sivarajkumar, Sonish**, and Yanshan Wang. "HealthPrompt: A Zero-shot Learning Paradigm for Clinical Natural Language Processing", *AMIA Annual Symposium (AMIA 2022), as one of 8 finalists in Best Paper competition 2022.*
- Oniani, David, **Sonish Sivarajkumar**, and Yanshan Wang. "Few-Shot Learning for Clinical Natural Language Processing Using Siamese Neural Networks",
  Journal of Medical Internet Research-Artificial Intelligence (JMIR-AI 2023)
- **Sivarajkumar S**, Tandale P, Bhardwaj A, Johnson KW, Titus A, Glicksberg BS, Khader S, Yadav KK, Subramanian L. Generation of a Compendium of Transcription Factor Cascades and Identification of Potential Therapeutic Targets using Graph Machine Learning. *Under Review(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 (2022)*
- 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, "Mitigating Bias in the Digital Twin Learned from Electronic Health Records", *Digital Health Summit 2023*
- **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.*

OTHER PROFESSIONAL ACTIVITIES

Editorial Activities

Journal of the American Medical Informatics Association (JAMIA)-Student Editorial Board Member | *2024 - present*

American Medical Informatics Association (AMIA) Newsletter- Junior Editor | *2022 - present*

PC Member

Clinical Natural Language Processing Workshop – NAACL | *2023*

Biomedical Natural Language Processing Workshop – ACL| *2023*

ICHI (IEEE International Conference on Healthcare Informatics) | *2023, 2024*

Peer Review

Journal of Medical Internet Research (JMIR) | *2025*

Plos One | *2024*

Journal of the American Medical Informatics Association (JAMIA) | *2024*

IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI) | *2023, 2024*

ACL (Association of Computational Linguistics) | *2023*

ICHI (IEEE International Conference on Healthcare Informatics) | *2023, 2024*

Journal of Healthcare Informatics Research (JHIR) |2023

ICHI (IEEE International Conference on Healthcare Informatics) | *2022*

ICML (International Conference on Machine Learning) | *2022*

LREC (Conference on Language Resources and Evaluation) | *2022*

Workshops

Publication chair EBAIC 2024 (International Workshop on Ethics and Bias of Artificial Intelligence in Clinical Applications) |*San Francisco, 2024*

Publication chair EBAIC 2023 (International Workshop on Ethics and Bias of Artificial Intelligence in Clinical Applications) |*Houston, 2023*

Co-organized 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

- Merck Inspire Awards | *2023*
- Fellowship – School of Computing and Information, University of Pittsburgh| *2021-2023*
- AMIA 2022 Best Paper Award Finalist | *2022*
- IQVIA Impact Award | *2021*