Suraj Rajendran

8360 Village School Rd., Cumming, Georgia 30332

https://github.com/surajraj99 sur4002@med.cornell.edu 225-587-0371

Education

Weill Cornell Medical College — **New York City, NY**
Tri-Institutional PhD Program in Computational Biology — *Expected Graduation: June 2025*

Georgia Institute of Technology — **Atlanta, GA**
BS in Biomedical Engineering GPA: 3.96/4.0 — *May 2021*

- Minor in Computing and Intelligence GPA: 4.0/4.0
- President's Undergraduate Research Award (PURA) Fellowship
- Faculty Honors

Research Experience

Weill Cornell Medical College — **New York City, NY**
Graduate Rotation Researcher — *August 2021 - Present*

Investigating Impact of Data Heterogeneity in Federated Learning

- Delineated a protocol to test the impact of data heterogeneity in local, pooled, and federated model settings
- Collected Acute Kidney Injury (AKI) and Sepsis data from 7 different hospitals within the New York medical system
- Adapted the SecureBoost algorithm to investigate the effect of data discrepancies within federated XGBoost
- Ranked 354 medications, lab test measurements, and vital signs, uniquely important at each New York hospital

Predicting Ploidy Status of Embryos Using Deep Video Classification

- Developed a pipeline to train models on in vitro fertilization time-lapse videos to determine ploidy status for 2000 embryos
- Collaborated with embryologists to determine ways to standardize videos of embryos to ensure non-biased predictions
- Created a pre-trained CNN-LSTM architecture based on ImageNet to process and predict ploidy status
- Explored state of the art video classification models such as 3DConvNet and I3D to determine efficacy of application

Wake Forest School of Medicine — **Winston-Salem, NC**
Bioinformatics Researcher, PI: Dr. Umit Topaloglu — *May 2019 - Present*

Privacy Preserving Methods Through Holographic Transformations

- Performed frequency domain transformations on multiple standard ML datasets to mask information more efficiently
- Created pipeline for frequency domain masking of datasets, allowing for information control by active parties
- Validated transformed dataset security through simulating a Generative Adversarial Network (GAN) attack

Predicting COVID-19 Diagnosis Using Televisit and Progress Notes

- Investigated the application of 3 deep neural network models on patient notes to extract COVID-19 symptoms
- Preprocessed and cleaned patient notes via Python's NLTK NLP, regex, and autocorrect python libraries
- Trained DNN models on notebook instance provided by Google Cloud's High-Performance Computing services
- Classified over 40,000 presumptive patients as COVID positive or negative using best performing DNN models

Federated Learning Using Cloud Computing

- Developed environment to test the efficiency and performance of three transfer learning methods
- Created a protocol for sharing models via GitHub from one institution to another without sharing private data
- Delineated the results of the developed federated learning mechanisms in a manuscript submitted to JCO CCI

Detecting Smoking Status Using Natural Language Processing

- Created an NLP Pipeline that cleans EHR data using Python libraries such as gensim, spaCy, and Keras
- Developed 6 machine learning models that classified patients based on their smoking status using progress notes
- Determined optimal parameters for learning rate, batch size, and loss functions using different search techniques

Professional Experience

Colgate-Palmolive — **Topeka, KS**
Research and Development Bioinformatics Intern — *June 2021 - Aug 2021*

Diagnosis Mapping and Searching within Veterinary Notes

- Developed two pipelines for standardizing new diagnoses entered by clinicians by using NLP and SNOMED relations
- Created an algorithm to find diagnoses within medical notes supplementing veterinarian diagnosis
- Added thresholds and features to the pipelines to allow for user modulation based on use application
- Validated results of the pipelines with clinicians and the ground truth to verify high performance and accuracy
- Presented pipelines and algorithms to professionals and created documentation to facilitate future use of the built programs
- Constructed a multimodal Long Short-Term Memory (LSTM) model for early prediction of renal disease

Traits AI **San Diego, CA**
Software Engineering Intern *August 2018 - September 2019*
Voice Assistants for Google Assistant and Amazon Alexa
- Coded script and UI which users would be interacting with on the Google Assistant Agents or Amazon Alexa
- Created 3 Voice Activated Conversational AI using Google SDK, Node.js, and the Jovo Database
- Monitored user interaction with Google Agents and Alexa Skills using Dialogflow Analytics and DynamoDB
- Managed and regulated the AWS Lambda servers on which Actions and Skills were hosted

Publications

- **Rajendran S**, Topaloglu U. *A Cloud Based Federated Learning Implementation Across Medical Centers.* JCO Clinical Cancer Informatics (2021)
- **Rajendran S**, Topaloglu U. *Extracting Smoking Status from Electronic Health Records Using NLP and Deep Learning.* AMIA Jt Summits Translational Sci Proc. (2020) 507-516
- **Rajendran S**, Ong T, Zameza P, et al. *Including social determinants of health in prognostic models for intracerebral hemorrhage.* Society of Critical Care Medicine. (2021)

Projects

Center of Disease Control Threat Detection Challenge *Decemeber 2021 - Present*
- Trained an LSTM network on textual input from social media to predict Adverse Childhood Experiences (ACEs)
- Performed cluster analysis to determine similarities between Reddit posts focused on traumatic childhood events
- Generated a graph network with 50+ nodes depicting Twitter users and their likelihood of having experienced an ACE
- Conducted social network analysis (SNA) to determine communities in which external intervention was necessary

Department of Education Automated Scoring Challenge *November 2021 - January 2022*
- Constructed 10+ machine learning models to score constructed response items for the NAEP's reading assessments
- Utilized BERT language models through Pytorch in combination with natural language cleaning processes for classification
- Fulfilled the performance requirements to use automated models in a real-world setting within 5% margin of error
- Accounted for racial and gender disparities within student data using data augmentation to minimize model bias

Predicting Criminal Recidivism Using Feature Engineering and XGBoost *June 2021 – October 2021*
- Utilized state of the art machine learning techniques to assist in predicting recidivism to aid in evaluating prison efficiency
- Preprocessed and standardized large datasets to ready them for a XGBoost model with fine-tuned parameters
- Added features to the dataset to ensure that models were not biased against certain demographics

Deriving Actionable Strategies Using Machine Learning *January 2021 – February 2021*
- Analyzed historical CarMax data to identify trends in customer purchases and preferences across many demographics
- Developed neural network and random forest models to predict customer decisions based on various attributes
- Determined marketing inventory strategies for CarMax to utilize to draw in distinct segments of customers
- Created an interface which allows personalized experience custom fit for each customer to maximizes their satisfaction

Fellowships and Accolades

Fellowships

National Science Foundation Graduate Research Fellowship *August 2022 - August 2025*
Role: PhD Student
Stipend: $34,000

Honors & Awards

Georgia Tech Hacklytics 2022 [2nd Place Healthcare Hack and 2nd Place Overall] *February 2022*
CDC & NASA Detecting Emerging Threats Challenge [Won 1st Place - $7500 Cash Prize] *January 2022*
DOE Automated Scoring Challenge [Won 4th Place - $1250 Cash Prize] *January 2022*
DOJ Recidivism Forecasting Challenge [Won 4 Awards totalling $23,000] *August 2021*
CarMax ML/AI Data Analytics Showcase [1st Place out of 200 Teams - $3000 Cash Prize] *February 2021*
AAMI Foundation's Michael J. Miller Scholarship [$3000 Cash Prize] *January 2021*

Skills

- **Software:** google cloud | keras (tensor flow backend) | pytorch | SQL | natural language processing
- **Programming:** Python | Java | MATLAB | Node.js | R | Github | bash
- **Communication:** statistics | technical writing
- **Languages:** English – native | French – intermediate
- **Certifications:** CITI Training | Certified SOLIDWORKS Professional