KAUSTUBH SAPRU

646 417 4838 |kaustubh.sapru@gmail.com |github.com/KaustubhSapru92 | LinkedIn

CAREER HIGHLIGHTS

AI Researcher with 3 years of experience in writing production-ready code to develop, deploy and scale AI/ML systems. Passion for extracting meaningful insights driven by feature engineering on complex datasets for better model learning backed with extensive statistical analyses and extensive data visualization. Working knowledge in developing machine learning pipelines from inception to completion in areas of Computer Vision, Natural Language Processing (NLP) and Multi-modal systems. Ideated on new ML approaches to drive rapid prototyping and production contracts. Highly motivated curiosity driven approach with ability to explore uncharted territory.

TECHNICAL SKILLS

- **Programming languages : Python, Matlab, C++, SQL, Tableau**
- **Frameworks : PyTorch, TensorFlow, Keras, Git, OpenCV, ScikitLearn, Pandas, OpenCV, Docker, IntelliJ**
- **Cloud Platforms : Google Cloud Platform, Microsoft Azure, AWS, Git Codespace**
- **Certifications : Big Data - MySQL, Deep Networks - PyTorch, AI Workflow, Time Series Prediction**

WORK EXPERIENCE

Computational Scientist *– Advanced Science Research Center/RF CUNY* *Sep 2022 – Sep 2023*

Predictive Modeling for Time Series Multi-Channel Experimental Signal

- Collaborated with stakeholders to identify opportunities for AI-driven solutions to drug discovery/manufacturing. Deployed scaled end-to-end standalone application that analyzes rare proteins through the NMRFx Analyst Software (**nmrfx.org/analyst**).
- Designed and trained bespoke multi-cascaded network architecture consisting of FNet Auto-Encoders, Time to Vector Transformers, Residual Convnets on multi-channel time domain experimental data to deliver predictions for properties of rare proteins.
- Utilized GPU cluster backed parallel processing through TensorFlow for accelerated training efficiency on training data and automated training, testing through Docker.
- Implemented evaluation metrics to test efficacy of deployed model and deployed statistical measures to compare model performance with other model architectures.

Generative AI for Predicting Protein Movement Patterns

- Explored generative AI models like GANs and Diffusion models for predicting movement patterns for protein structures via self-supervised learning to improve drug discovery.
- Created a multi-modal generator architecture to process signals and NMR data recorded at different magnetic field strengths to estimate protein movement.
- Collaborated with interdisciplinary teams to employ bayesian weighted sampling of training data taken at different field strengths to ensure scalability and higher prediction accuracy.

Conference: Experimental NMR Conference 2023 (Poster), NY Structural Biology Discussion Group Conference

Data Scientist *– Center for Discovery and Innovation(Neural Engineering Group)* *Jun 2021 – Sep 2022*

Computer Vision Software to Analyze Neural Response to Cartoon Faces in Videos

- Collaborated with cross functional teams to develop computer vision software to track, curate and process eye movements of human subjects to face movements of characters in animated movie clips.
- Spearheaded the strategy to track face movements of specific characters throughout movie clips using deep learning object detection networks such as Detectron2(by Facebook) on 36000 frames.
- Developed an application to automate identification of inaccurate face labels and their replacement with correct ones. Thesis : **https://academicworks.cuny.edu/cc_etds_theses/1066/**
- Engineered training data to improve baseline bounding box average precision of the model from 43% to a maximum of 71% and a classification accuracy of 80%. **Code : https://lnkd.in/dB2Z6vJ3**
- Transformed eye tracking data into machine-readable features, used machine learning to select most important features, training a classifier to classify 19,565 data points as going to face or away from face.

Predictive Modeling to Detect Anomalies in Stock Market Data

- Developed a model to predict a halt in stock market trading precisely one minute before its onset.
- Analyzed multiple days of stock trading to generate insights on trading patterns to devise a strategy to understand dynamics around a stock market halt (circuit breaker event).
- Deployed machine learning to establish the rate of price drop within each minute of the trading. Executed the trained model on data for 20th March 2020 to accurately predict the circuit breaker event for the day and to test its efficacy. **Code : https://lnkd.in/dJaM88TC**

Impact of COVID-19 on Education: A Machine Learning and Statistical Analysis

- Organized and conducted a study on a group of 150 students to determine which group based on gender, age and education level would prefer online education to in-person education.
- Performed statistical testing to determine that less than 50% students prefer online education with 99% confidence($p<0.01$).
- Employed machine learning to determine probabilities of groups, based on gender, age and education, to prefer online education to in-person education. Two groups emerged with the highest probability: 1) Male, Age 25-43 with a master's degree and 2) Female, Age: 18-24 having a bachelor's degree.
- Transformed the data to achieve accuracy of 86.3% compared to 69% on raw data revealing a 90.6% area under the ROC curve. **Code : https://lnkd.in/eQfMvCW**

Conferences: Neuroergonomics Conference and NYC Neuromodulation Conference 2022

Research Scientist – *Chimera Translational Research Fraternity Pvt Ltd.* *Feb 2019 – Apr 2019*

- Managed and expanded the Human Leukocyte Antigen (HLA) matchmaking database used for matching patients with blood disorders through stem cell transplants with possible suitable donors.
- Set up stem cell donation camps to attract future possible donors. Stored sequenced data in the matchmaking database for the donor's chromosome 6, exon 1, exon 2 to determine HLA type (High resolution HLA typing).
- Utilized machine learning to identify best possible patient-donor matches for 14 HLA types highly prevalent in the Indian population.
- Submitted a proposal to indigenously develop a HLA matching kit specifically for the 14 identified HLA types. (HLA matching kit is imported)

Junior Research Fellow – *Indian Institute of Science, Bangalore (IISc)* *Aug 2016 - Jan 2019*

Computer Aided Classification of Different Stages of Breast Cancer

- Identified ICIAR2018 breast cancer histology dataset with 400 images attributed to four classes namely: normal, benign, in situ carcinoma, invasive carcinoma
- Trained a convolutional neural net classifier using Pytorch, accuracy = 80%

Project Associate – *Indian Institute of Technology, Delhi* *Aug 2015 - Nov 2015*

Image Processing Tool for Non-Invasive Glucose Detection

- Developed an open-source image processing tool to analyze optical biosensor data to determine glucose concentration in a salivary sample
- Implemented the algorithm using Android Studio and developed an android application

ACADEMIC AWARDS AND HONORS

- Rukin Award For Academic & Professional Perseverance ($2000) - Biomedical Engineering(BME), CCNY
- Neural Engineering Group Scholarship ($2000)
- All India Rank 62 (99.4 percentile), Graduate Aptitude Test of Engineering (GATE 2016)
- Thapar University Scholarship 2012 (27,500 Indian Rupees)

EDUCATION

MS, Biomedical Engineering – The City College New York, CUNY (GPA : 4.0) *Aug 2020 - Sep 2022*
Post Graduate Diploma, ML & AI – Amity University Online (GPA : 8.35/10) *Jul 2019 - Jul 2020*
Bachelor of Technology, Biotechnology – Thapar University (GPA : 6.81/10) *Aug 2011 - Jun 2015*

PUBLICATIONS

Nentwich M, Leszczynski M, Russ BE, Hirsch L, Markowitz N, **Sapru K**, Schroeder CE, Mehta AD, Bickel S, Parra LC. Semantic novelty modulates neural responses to visual change across the human brain. Nat Commun. 2023 May 22;14(1):2910. doi: 10.1038/s41467-023-38576-5. PMID: 37217478; PMCID: PMC10203305.

TEACHING ASSISTANTSHIP

- TA - Image and Signal Processing, BME CCNY (Spring 2022), Instructor : Prof. Lucas Parra
- CUNY Tutor Corps - Computer Science and Computer Programming (Fall 2021)

OTHER EXPERIENCE / VOLUNTEER WORK

- Research Internship 2014 with Prof. Sheila MacNeil - Kroto Research Institute, University of Sheffield, UK
- Traveled to Cold Spring Harbor Laboratory during June - July 2023 for interactive sessions on Machine Learning for STAMPR bottlenecks.
  Volunteer Work :
- High School Exam Writing for Blind Students - 2010
- English, Math Classes for UnderPrivileged High/Middle School Students - 2023

EXTRA CURRICULAR ACTIVITIES

- Proficient Guitar player/Singer with a Youtube channel/blog which has accumulated views of over two million. https://www.youtube.com/user/metalgodmaiden
- Performed (with my band Prithvi) and part of the organizing team of Kashmiri Pandit Youth Festival 2011.
- Headlined SHUHUL TAAP III, a 3 day Kashmiri Pandit Cultural event, presided by Chief Minister of Delhi, Smt. Sheila Dixit and various renowned clubs such as Hard Rock Café, Café Morrison, Turquoise Cottage etc with my band Prithvi.
- Convener of the MUDRA (Music & Dramatics) Society of Thapar University. Led organizing committee of stage performances and all cultural events in the University.
- Organizing team of Annual University Cultural Night 2012, 2013 and SUR (Intra- College Musical event) 2011,2012,2013,2014.
- First prize in Instrumental competition, IZHAAR 2012 (Intra-College Cultural competition).
- First prize in Punjabi Solo, Punjabi Duet in IZHAAR 2013, Hindi Duet in IZHAAR 2014
- Player of University Cricket Team, Winner of Inter-Engineering; Inter University Cricket Tournament 2015, held at Punjab Engineering College (PEC).

KAUSTUBH SAPRU

646 417 4838 |kaustubh.sapru@gmail.com |github.com/KaustubhSapru92 | LinkedIn

CAREER HIGHLIGHTS

AI Researcher with 3 years of experience in writing production-ready code to develop, deploy and scale AI/ML systems. Passion for extracting meaningful insights driven by feature engineering on complex datasets for better model learning backed with extensive statistical analyses and extensive data visualization. Working knowledge in developing machine learning pipelines from inception to completion in areas of Computer Vision, Natural Language Processing (NLP) and Multi-modal systems. Ideated on new ML approaches to drive rapid prototyping and production contracts. Highly motivated curiosity driven approach with ability to explore uncharted territory.

TECHNICAL SKILLS

- **Programming languages : Python, Matlab, C++, SQL, Tableau**
- **Frameworks : PyTorch, TensorFlow, Keras, Git, OpenCV, ScikitLearn, Pandas, OpenCV, Docker, IntelliJ**
- **Cloud Platforms : Google Cloud Platform, Microsoft Azure, AWS, Git Codespace**
- **Certifications : Big Data - MySQL, Deep Networks - PyTorch, AI Workflow, Time Series Prediction**

WORK EXPERIENCE

Computational Scientist *– Advanced Science Research Center/RF CUNY* *Sep 2022 – Sep 2023*

Predictive Modeling for Time Series Multi-Channel Experimental Signal

- Collaborated with stakeholders to identify opportunities for AI-driven solutions to drug discovery/manufacturing. Deployed scaled end-to-end standalone application that analyzes rare proteins through the NMRFx Analyst Software (**nmrfx.org/analyst**).
- Designed and trained bespoke multi-cascaded network architecture consisting of FNet Auto-Encoders, Time to Vector Transformers, Residual Convnets on multi-channel time domain experimental data to deliver predictions for properties of rare proteins.
- Utilized GPU cluster backed parallel processing through TensorFlow for accelerated training efficiency on training data and automated training, testing through Docker.
- Implemented evaluation metrics to test efficacy of deployed model and deployed statistical measures to compare model performance with other model architectures.

Generative AI for Predicting Protein Movement Patterns

- Explored generative AI models like GANs and Diffusion models for predicting movement patterns for protein structures via self-supervised learning to improve drug discovery.
- Created a multi-modal generator architecture to process signals and NMR data recorded at different magnetic field strengths to estimate protein movement.
- Collaborated with interdisciplinary teams to employ bayesian weighted sampling of training data taken at different field strengths to ensure scalability and higher prediction accuracy.

Conference: Experimental NMR Conference 2023 (Poster), NY Structural Biology Discussion Group Conference

Data Scientist *– Center for Discovery and Innovation(Neural Engineering Group)* *Jun 2021 – Sep 2022*

Computer Vision Software to Analyze Neural Response to Cartoon Faces in Videos

- Collaborated with cross functional teams to develop computer vision software to track, curate and process eye movements of human subjects to face movements of characters in animated movie clips.
- Spearheaded the strategy to track face movements of specific characters throughout movie clips using deep learning object detection networks such as Detectron2(by Facebook) on 36000 frames.
- Developed an application to automate identification of inaccurate face labels and their replacement with correct ones.
- Engineered training data to improve baseline bounding box average precision of the model from 43% to a maximum of 71% and a classification accuracy of 80%. **Code : https://lnkd.in/dB2Z6vJ3**
- Transformed eye tracking data into machine-readable features, used machine learning to select most important features, training a classifier to classify 19,565 data points as going to face or away from face.

Predictive Modeling to Detect Anomalies in Stock Market Data

- Developed a model to predict a halt in stock market trading precisely one minute before its onset.
- Analyzed multiple days of stock trading to generate insights on trading patterns to devise a strategy to understand dynamics around a stock market halt (circuit breaker event).
- Deployed machine learning to establish the rate of price drop within each minute of the trading. Executed the trained model on data for 20th March 2020 to accurately predict the circuit breaker event for the day and to test its efficacy. **Code : https://lnkd.in/dJaM88TC**

Impact of COVID-19 on Education: A Machine Learning and Statistical Analysis

- Organized and conducted a study on a group of 150 students to determine which group based on gender, age and education level would prefer online education to in-person education.
- Performed statistical testing to determine that less than 50% students prefer online education with 99% confidence($p<0.01$).
- Employed machine learning to determine probabilities of groups, based on gender, age and education, to prefer online education to in-person education. Two groups emerged with the highest probability: 1) Male, Age 25-43 with a master's degree and 2) Female, Age: 18-24 having a bachelor's degree.
- Transformed the data to achieve accuracy of 86.3% compared to 69% on raw data revealing a 90.6% area under the ROC curve. **Code : https://lnkd.in/eQfMvCW**

Conferences: Neuroergonomics Conference and NYC Neuromodulation Conference 2022

Junior Research Fellow *– Indian Institute of Science, Bangalore (IISc)* *Aug 2016 - Jan 2019*

Computer Aided Classification of Different Stages of Breast Cancer

- Identified ICIAR2018 breast cancer histology dataset with 400 images attributed to four classes namely: normal, benign, in situ carcinoma, invasive carcinoma
- Trained a convolutional neural net classifier using Pytorch, accuracy = 80%

Project Associate *– Indian Institute of Technology, Delhi* *Aug 2015 - Nov 2015*

Image Processing Tool for Non-Invasive Glucose Detection

- Developed an open-source image processing tool to analyze optical biosensor data to determine glucose concentration in a salivary sample
- Implemented the algorithm using Android Studio and developed an android application

ACADEMIC AWARDS AND HONORS

- Rukin Award For Academic & Professional Perseverance ($2000) - Biomedical Engineering(BME), CCNY
- Neural Engineering Group Scholarship ($2000)
- All India Rank 62 (99.4 percentile), Graduate Aptitude Test of Engineering (GATE 2016)
- Thapar University Scholarship 2012 (27,500 Indian Rupees)

EDUCATION

MS, Biomedical Engineering – The City College New York, CUNY (GPA : 4.0) *Aug 2020 - Sep 2022*

Post Graduate Diploma, ML & AI – Amity University Online (GPA : 8.35/10) *Jul 2019 - Jul 2020*

Bachelor of Technology, Biotechnology – Thapar University (GPA : 6.81/10) *Aug 2011 - Jun 2015*

PUBLICATIONS

Nentwich M, Leszczynski M, Russ BE, Hirsch L, Markowitz N, **Sapru K**, Schroeder CE, Mehta AD, Bickel S, Parra LC. Semantic novelty modulates neural responses to visual change across the human brain. Nat Commun. 2023 May 22;14(1):2910. doi: 10.1038/s41467-023-38576-5. PMID: 37217478; PMCID: PMC10203305.

TEACHING ASSISTANTSHIP

- TA - Image and Signal Processing, BME CCNY (Spring 2022), Instructor : Prof. Lucas Parra
- CUNY Tutor Corps - Computer Science and Computer Programming (Fall 2021)

EXTRA CURRICULAR ACTIVITIES

- Proficient Guitar player/Singer with a Youtube channel/blog which has accumulated views of over two million. https://www.youtube.com/user/metalgodmaiden
- Performed (with my band Prithvi) and part of the organizing team of Kashmiri Pandit Youth Festival 2011.
- Headlined SHUHUL TAAP III, a 3 day Kashmiri Pandit Cultural event, presided by Chief Minister of Delhi, Smt. Sheila Dixit and various renowned clubs such as Hard Rock Café, Café Morrison, Turquoise Cottage etc with my band Prithvi.
- Convener of the MUDRA (Music & Dramatics) Society of Thapar University. Led organizing committee of stage performances and all cultural events in the University.
- Organizing team of Annual University Cultural Night 2012, 2013 and SUR (Intra- College Musical event) 2011,2012,2013,2014.
- First prize in Instrumental competition, IZHAAR 2012 (Intra-College Cultural competition).
- First prize in Punjabi Solo, Punjabi Duet in IZHAAR 2013, Hindi Duet in IZHAAR 2014
- Player of University Cricket Team, Winner of Inter-Engineering; Inter University Cricket Tournament 2015, held at Punjab Engineering College (PEC).