MANAN SAXENA

State College, PA, 16801 (Willing to relocate) | 814.769.0852 | manansaxena05@gmail.com |LinkedIn | Website

EDUCATION

Doctor of Philosophy, Informatics
Penn State University | State College, PA | GPA: 4.0 — Jan 2025 – May 2027

Masters of Science, Informatics
Penn State University | State College, PA | GPA: 3.92 — Aug 2022 – May 2024

Bachelor of Technology, Software Engineering
Delhi Technological University | Delhi, India | GPA: 3.7 — Aug 2016 – May 2020

SKILLS

Programming Languages
C++, Python, R, SQL, JavaScript, Stan

Development Tools
Google Cloud Platform, Databricks, jQuery, Docker, GitHub, Linux, Jenkins

Data Science and AI Tools
PyTorch, TensorFlow, Keras, LangChain, MLflow, Streamlit, PyMC, Bambi, Prophet, Scikit-Learn, OpenCV, Tidyverse, Tableau

WORK EXPERIENCE

Data Scientist (Internship)
Hartford Steam Boiler (Munich Re) | Hartford, CT
Supervisor: Dr. Yue Tang — May 2025 – Jul 2025

- Developed reserve recommendation models for equipment breakdown insurance claims, focusing on claim severity prediction
- Engineered and compared advanced models (XGBoostLSS, GAM, GAMLSS), improving NRMSE from **0.93 to 0.86**
- Enhanced claim notes feature extraction using Llama 3 with prompt engineering, boosting **accuracy from 51% to 87%**
- Reduced LLM processing time for feature extraction from **40 hours to 15 hours for ~21k samples** via parallelized batch inference
- Prototyped Bayesian hierarchical models for claim severity using Bambi & PyMC, capturing the true data-generating process; integrated with MLflow for experiment tracking
- Worked with Databricks and Spark to handle large-scale insurance datasets efficiently

Software Engineer
Tummee.com | Remote — Sep 2021 – Jun 2022

- Managed end-to-end software development of a feature addition to the core sequence builder functionality of the platform
  - Ensured **0 fault** live deployments and optimal cross-platform performance
  - Secured over **800 users within a month** of release
- Led a cross-functional team to revamp the customer issue submission portal, leveraging customer insights to improve user experience, resulting in a **20% reduction in issue resolution time**
- Developed REST-APIs in Python's webapp2 framework integrated with GCP. Built UIs using Bootstrap and handled dynamic behavior with JavaScript and jQuery
- Automated the transformation of unstructured data to the structured database using Google Sheets API, **reducing manual data entry time by 50%**

RESEARCH EXPERIENCE

Graduate Research Assistant

Pennsylvania State University | State College, PA — May 2023 – Present

Supervisor: Dr. Justin Silverman

- Developed a scalable Bayesian inference algorithm for multivariate count time series data, applied to understanding trends and patterns in microbial systems
- Calculated closed-form gradients for posterior estimation, achieving **20-30x faster optimization** than automatic differentiation in Stan, and generated 95% credible intervals using Multinomial Dirichlet Bootstrap with **almost 0 deviation** from the true posterior
- Created an R package called Fenrir with base code in C++ utilizing Eigen and Boost libraries for optimized performance [Link]
- Engineered a codebase employing shell scripts to run automated jobs with minimal user input on Penn State's High-Performance Compute (HPC) [Link]

Machine Learning Researcher

Trinity College Dublin | Dublin, Ireland — Jun 2019 – Jul 2021

Supervisor: Dr. Ciaran Simms, Dr. Aljosa Smolic, Dr. Richard Blythman

- Developed an automated end-to-end pipeline using fine-tuning of deep learning models for predictive analytics in sports injury prevention. Used 3D pose estimation, object tracking, and instance image segmentation models
- Tested proof-of-concept level model on novel rugby tackle data set, comparing to industry benchmark motion capture systems (VICON) with reasonable performance and at a fraction of the cost
- Built an automated pipeline for camera calibration and face blurring to acquire rugby tackle datasets [Link] [Link]
- Collaborated with coaches and physiotherapists to translate domain knowledge into decisions for prototype development, resulting in more effective injury prevention strategies

TALKS

Causal Representation Learning — Jul 2025

- Paper Discussion, Data Science Team, Hartford Steam Boiler

Scalable Inference for Bayesian Multinomial Logistic-Normal Dynamic Linear Models — Oct 2024

- Bioinformatics Method Developers Community Day, Center for Computational Biology and Bioinformatics, Pennsylvania State University

PUBLICATIONS

- Saxena, M., Chen, T., & Silverman, J. D. (2025). Scalable inference for Bayesian multinomial logistic-normal dynamic linear models. Accepted in 28th International conference on artificial intelligence and statistics (AISTATS) [Link]
- Blythman, R., Saxena, M., Tierney, G. J., Richter, C., Smolic, A., & Simms, C. (2022). Assessment of deep learning pose estimates for sports collision tracking. Journal of sports sciences, 40(17), 1885-1900 [Link]
- Dhiman, C., Saxena, M., & Vishwakarma, D. K. (2019, September). Skeleton-based view invariant deep features for human activity recognition. In 2019 IEEE fifth international conference on multimedia big data (BigMM) (pp. 225-230). IEEE. [Link]
- Garg, A.*, Aggarwal, K.*, Saxena, M.*, & Bhat, A. (2021). Classifying medical histology images using computationally efficient CNNs through distilling knowledge. In Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2020, Volume 3 (pp. 713-721). Springer Singapore [Link]

CERTIFICATIONS AND AWARDS

2024	Databricks Fundamentals [Link]
2016	All India Rank 2,661 out of 1.4 million candidates in Joint Entrance Examination Mains (JEE Mains)
2012	National Talent Search Examination scholar, ranked among the top 1,000 out of 1.2 million candidates, was awarded a scholarship until the completion of undergraduate studies.

PROJECTS

Baseball: Modeling Batter's Swing Probability Jun 2024

- Developed a predictive model for batter's swing probability for different types of pitches using a Random Forest classifier and Grid Search Cross Validation, **achieving an accuracy score of 89%**
- Introduced the Swing Efficiency Index (SEI) metric, which measures the ratio of actual swing percentages to the adjusted swing probability based on pitch height, providing insights into the batter's mentality and execution. Additionally, analyzed middle-middle pitches and presented findings in a format understandable by coaches [Link]

Forecasting SARS-CoV-2 Concentrations in Wastewater May 2024

- Created and deployed a time-series forecasting model for SARS-CoV-2 Concentrations using Prophet integrated with MLflow for tracking, improving RMSE from **140 in the trend model to 125**. Built containerization and CI/CD pipeline development using Docker and Jenkins, facilitating deployments to GitHub Container Registry [Link]

Music Lyrics Analysis and Q&A System Feb 2024

- Deployed an interactive system on Streamlit for analyzing and responding to queries about music lyrics. Utilized LangChain framework, combined with ChatGPT (Large Language Model) API for natural language processing (NLP) and YouTube API for lyric extraction from music videos [Link]

Classification of Breast Cancer Histology Images through Distilling Knowledge Aug 2019 – Jun 2020

- Implemented a light CNN model for high-resolution breast cancer histology image classification, utilizing knowledge distillation techniques and attention maps. Leveraged ResNet 50 as a teacher model to improve the performance of a lighter ResNet 8 model, boosting its **accuracy from 75% to 80%**

Skeleton-Based View Invariant Deep Features for Human Activity Recognition Dec 2018 – May 2019

- Introduced novel view-invariant skeletal features to describe spatial-temporal characteristics of human motion. Achieved a **2% accuracy improvement** over existing state-of-the-art models on the NUCLA dataset through the application of transfer learning and dynamic image techniques