Resume of Sohel Ahmed

Phone: 607-262-3293
Email: sohelcu06@gmail.com
Address: 12007 N. Lamar Blvd, TX 78753

Professional Summary

- Experienced Research Scientist with over 8 years in data science and analytics, specializing in Generative AI and machine learning. Proven expertise in designing and implementing data-driven solutions across diverse fields, including public health and logistics. Adept at survey design, data visualization, and stakeholder management with a strong background in research methodologies and employee well-being metrics. Proven leader in managing and executing complex projects from inception to completion.

PROFESSIONAL EXPERIENCE

- **Research Scientist, Texas A&M Transportation Institute, TX 78752** **Oct 2023 – Present**

  - **Project listing Automation Generative Artificial Intelligence Architect (Python, Flask, HTML, SQL)**
    - Developed an end-to-end **Medical Chatbot** using Python, LangChain, GPT, Pinecone, and AWS CICD for deployment.
    - Automated the deployment pipeline with GitHub Actions, Docker, ECR, and EC2 to efficiently manage and deploy the chatbot's containerized environment on AWS, supporting conversational interactions through OpenAI's GPT and Pinecone's vector database.
    - Developed a Source-Code-Analysis Generative AI application using Python, LangChain, Flask, and OpenAI's GPT-3, with ChromaDB for efficient data retrieval, enabling analysis and interaction with source code via a locally hosted interface.
    - Built a **Retrieval-Augmented Generation (RAG) App** using Streamlit and Langchain, integrating OpenAI's GPT to answer user queries by retrieving and processing documents from specified URLs, enhancing document-based Q&A functionality with similarity search and a user-friendly interface.
    - Developed **LLM_Apps_Chainlit**, a demo application using Chainlit to build and deploy LLM-based apps with real-time conversational interactions, leveraging custom logic and efficient LLM integration.
    - Implemented automation techniques utilizing Access database, SQL, and Python, enhancing efficiency and accuracy in data management processes.
    - Developed and maintained a comprehensive project call database for the Unified Transportation Program (UTP) initiative and collaborated with cross-functional teams to identify database requirements.
  - **Project Call(Budget planning) for TxDot (Python, Access, SQL)**
    - Automated the allocation of budget across districts by developing a database system.
    - Transformed unstructured data into a structured format for enhanced data analysis and reporting.
    - Designed and implemented a comprehensive dashboard using SQL, Access, and Python to facilitate real-time data visualization and decision-making.
  - **Historical data Budget TxDot (Python, Access, SQL)**
    - Developed a Snowflake database for TxDOT, optimizing data storage and management for enhanced accessibility.
    - Designed an interactive **Tableau live spatial dashboard** showcasing 30 years of historical data (2000-2025) for UTP, utilizing map visualizations to analyze and compare regional data by district, providing actionable insights for stakeholders
    - Implemented methods for efficient data extraction from the data warehouse, facilitating timely and accurate data analysis for decision-making processes.

- **Postdoctoral Associate, Cornell University, Cornell Wildlife Health, NY 14850** **Jul 2021 – Oct 2023**

  - Analyzed data using data visualization tools and reported key features using statistic tools and supervised machine learning techniques to find significant features.
  - Defined Data Pipeline (ingest / clean / munge / transform) for feature extraction toward downstream classification.
  - Compiled data from various sources public and private databases to perform complex analysis and data manipulation for actionable result.
  - Analyzed and modeled structured data using advanced statistical methods and implemented algorithms and software needed to perform analyses.
  - Worked with data compliance teams, data governance teams and data collection teams to establish and maintain data models, metadata, data dictionaries, source fields and their definitions.
  - **Major Projects Completed**:
    - **Predicting Chronic Wasting Disease and SARS-Cov-2 in county Scale using machine learning (aws python, SQL, Dash, Stremlit)**
      - Developed ML classification algorithms for non-linear data to identify the top risk factors contributing to County CWD positivity, achieving high model accuracies (Random Forest: 92.12%, Decision Tree: 87.04%, Gradient Boosting: 85.36%, Light Gradient Boosting: 92.50%)
      - Executed the Travel Salesman Problem to optimize the route (2123 miles) for deer sample collection (115 locations) using Google OR tools and Gurobi Package, lead the project
    - **SARS-CoV-2 Infection in deer in New York State I (aws Python, SaTScan)**
      - Classified 7 hotspots among 693 cluster points in infection areas, identifying severe prevalence of the disease in White-tailed Deer (WTD) in NYS
      - Conducted logistic regression analysis, revealing higher infection rates in male WTD (OR=1.952) with a model accuracy of 90%
      - ML algorithms identified infected deer reasons in NY (LGBM: 76.50% accuracy)
    - **Predicted medical test results based on the medical claim data (aws Python, SQL, Streamlit)**
      - Applied ML models such as Random Forest and Logistic regression to classify the result
      - Deploy the Web app to display the predict result in real time
    - **Waterfowl Contamination in Atlantic Flyway (R, QGIS)**
      - Proposed the Kruskal-Wallis Nonparametric test, higher mercury (Hg) levels in female waterfowl ($p<0.005$)
      - Performed Tukey post hoc test, identifying higher TCDD levels in Mallard and American Black Duck
    - **Natural Language Processing (NLP) Sentiment data I Python, NLTK) Dash, Stremlit**
      - Analyzed Twitter text data using NLTK to uncover trends, sentiments, and patterns
    - **Tumor Detection and diabetes prediction I Python, Dash, Stremlit**
      - K-nearest neighbors (KNN) algorithm uses to classify tumors in human tissue data by their gene expression.
      - Identifying Significant Features or Genes for Tumor Detection using Decision Trees
    - **Generative AI in Brest Cancer Detection by GANs for Image generation I Python, Dash, Stremlit**
      - Through the GANs Discriminator training generate synthetic data through the generator to identify fake image of breast cancer.
      - Generator iterate improve synthetic images and can increase deceive the discriminator

Research &Teaching Assistant

Machine Learning in Data Science through the Python at Cornell Wildlife Health Lab **2023**
Advisor: Asst. Professor Krysten Schuler

Statistical Application in Data Science in R at Cornell Wildlife Health Lab **2021**
Advisor: Asst. Professor Krysten Schuler

Stat 114: Probability, Random Variables, and Random Processes **2017**
Advisor: Professor Hiroshi Morita

- **National Institute of Informatics (Researcher) | Tokyo, Japan** **Apr 2019 - Jun 2021**

  - Successfully detects 98% of mutants of 142 lines of Lasso code efficiently by R implementation.

- Tested classification Machine Learning Algorithms and performed Cross Validation for hyper-parameter tuning to optimize the models for unseen data.
- Forecasted Kernel based polyhedral region for testing ML code using Python (HSIC-LASSO)
- Implemented machine learning algorithms such as lasso regression, linear regression, random forest, KNN, SVM, neural network, and k-means.

EDUCATION

Ph.D. in Information Science and Technology **2015 - 2018**
Osaka University, Osaka, Japan
(Received Journal Yearly Best Paper Award from Sustainability Journal for a paper based on the Ph.D. work)

Masters in Statistics **2011-2012**
University of Chittagong, Chittagong, Bangladesh

Bachelor in Statistics **2006-2011**
University of Chittagong, Chittagong, Bangladesh

MAJOR SKILLS/EXPERTISE

- Strong analytical and quantitative background specialized in predictive analytics, machine learning, deep learning.
- Proficient in big data technologies and programming.
- More than 8 years of experience in developing machine learning based algorithms.
- Expertise in analytical models, designing algorithms, building models, developing data mining and reporting solutions that scale across a massive volume of structured and unstructured data.
- Developed more than 15 End to End Machine learning projects with deployment.
- Strong geospatial data handling background include build Machine learning models, and build Choropleth Map
- More than 3 years' experience of developing text and language data through the Natural language Processing
- Having experienced of large language model through the OpenAI, LangChain, Chainlit package in Python

TECHNICAL SKILLS

- ***Data Science*** (8 Years): Data Analysis, Predictive / Prescriptive Analytics, Supervised / Unsupervised
- ***Statistical Methods*** (8 Years): Random Forest, Decision Tree, Logistic Regression, Linear Regression, Machine Learning, Predictive Modeling, Prescriptive Modeling
- ***Programming/Scripting***: Python (5 years) [Jupyter, PyCharm, Visual Studio], R Programming (8 Years), SAS (2 Years), Flask (2 years), HTML (2 years)
- **Artificial Intelligence**: Fully connected Neural Networks, Convolutional Neural Networks, Recurrent Neural Networks, Retrieval Augmented Generation (RAG), Lang chain, OpenAI, Pinecone, Huggingface
- **Libraries**: Xgboost, TensorFlow, PyTorch, TidyVerse, Numpy, Scipy, Folium
- **Data Visualization**: Matplotlib, Ggplot2, Tableau, Plotly, QGIS
- **Data Queries**: SQL, PostgreSQL and Snowflake
- **Data Science Deliverables**: Actionable Insights, Enhanced Decision-Making, Real-Time Predictions / Forecasts, Automation with AI
- **Dev Ops**: Spark, AWS, Stremlit, AWS CICD, EC2, First API

Funding Secured:
Internal Grants

1. College of Veterinary Medicine: Research Grants Program in Animal Health
   Project title: A Spatial Analysis of SARS-Cov-2 hotspots and Hierarchical model in White-tailed Deer in New York State 2022-2023, Amount Awarded: $50000

External Grants

2. Center For Rural and Primary Healthcare (CRPH) Research Grants: $100000

Project Title: Machine Learning Analysis of Infant Mortality Rates by Zip Code in South Carolina both of Urban and Rural Areas

OTHER EXPERIENCE AND SKILLS

- Successfully supervised staff and graduate students, lead projects.
- Completed the Leadership Development Program by Cornell University 2021.

Oral Presentation

Sohel Ahmed, ..Krysten Schuler "Evaluation of SARS-CoV-2 infection in white-tailed deer in New York State 2020-2021" Cornell Center for Immunology (CIHMID) Research Symposium 2022, Cornell University, Ithaca, NY

Sohel Ahmed, Fuyuki Ishikawa, and Mahito Sugiyama "Testing Machine Learning Code using Polyhedral Region" European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE-2020).

Publications

- **Ahmed, M.S.,** Hanley, B.J., Mitchell, C.I. *et al.* Predicting chronic wasting disease in white-tailed deer at the county scale using machine learning. *Sci Rep* **14**, 14373 (2024). https://doi.org/10.1038/s41598-024-65002-7

  Scientific Reports

  Altmetric Score: 53

  Social Media Attention: 8 tweeters

  Media Coverage: 8 news outlets

  Mendeley Readers: 3

  Impact: Ranked in the 96th percentile (11,592nd of 291,276 tracked articles of a similar age in all journals)

  Scientific Reports Ranking: 95th percentile (114th of 2,659 tracked articles of a similar age)

- Booth, J.G., Hanley, B.J., Hodel, F.H., Jennelle, C.S., Guinness, J., Them, C.E., Mitchell, C.I., **Ahmed, M.S**. and Schuler, K.L., 2023. Sample Size for Estimating Disease Prevalence in Free-Ranging Wildlife Populations: A Bayesian Modeling Approach. *Journal of Agricultural, Biological and Environmental Statistics*, pp.1-17.

- **Ahmed, M.S.,** Hanley, B.J., Mitchell, C.I. *et al.* Forecasting Chronic Wasting Disease (CWD) in Wisconsin State up to 2030 Using LSTM and ARIMA Models. *Submitted Paper*

- **Ahmed, M.S.,** Hanley, B.J., Mitchell, C.I. *et al.* Optimizing Deer Sample Collection Routes in Tennessee Using Vehicle Routing Problem Solutions and Web Application Development. *Submitted Paper*

- Real-time quaking-induced conversion (RT-QuIC) is robust for chronic wasting disease detection in white-tailed deer retropharyngeal lymph nodes: Joseph R.D; Alyssa W. K; **Md Sohel Ahmed,** Srinand Sreevatsan *(Submitted Paper*)

- Brenda J. Hanley, Michelle Carstensen, Daniel P. Walsh, Sonja A. Christensen, Daniel J. Storm, James G. Booth, Joseph Guinness, Cara E. Them, **Md Sohel Ahmed**, Krysten L. Schuler, Informing Surveillance through the

Characterization of Outbreak Potential of Chronic Wasting Disease in White-Tailed Deer, Ecological Modelling, Volume 471, 2022, 110054, ISSN 0304-3800,

https://doi.org/10.1016/j.ecolmodel.2022.110054.

- Hanley, B. J., Them, C. E., Mitchell, C. I., Carstensen, M., Walsh, D. P., Christensen, S. A., Storm, D. J., Booth, J. G., Guinness, J., Abbott, R. C., **Ahmed, M. S**., & Schuler, K. 2022. SLEI Model Software [Software]. Cornell University Library eCommons Repository. https://doi.org/10.7298/csew-h225.2

- Hanley, B., Mitchell, C. I., Abbott, R. C., Hollingshead, N., Carstensen, M., Walsh, D., Christensen, S., Storm, D., Kelly, J., Them, C., **Ahmed, M. S.,** Miller, L., & Schuler, K. 2021. Regional Wildlife Disease Hazard Software [Software]. Cornell University Library eCommons Repository. https://doi.org/10.7298/hqg5-ac08

- Brenda J. Hanley, Cara E. Them, **Md Sohel Ahmed**, Corey I. Mitchell, W. David Walter, Daniel P. Walsh, Christopher Jennelle, Nick A. Hollingshead, Rachel C. Abbott, James D. Kelly, Daniel M. Grove, Sonja Christensen, David Williams, Krysten L. Schuler Artificial Intelligence to Allocate Surveillance Resources for the Detection of Chronic Wasting Disease in Free-Ranging White-Tailed Deer (Working paper 2022)

- **Md Sohel Ahmed**, Cara E. Them b, James G. Booth c, Joseph Guinness d, Daniel P. Walshe, Brenda J. Hanley f, Krysten L. Schulerg Influential drivers of the reproduction number in chronic wasting disease in white-tailed deer(Progress Paper 2022).

- **Md Sohel Ahmed**, F. Ishikawa, M. Sugiyama (2020) Testing Machine Learning Code Using of Polyhedral region. In Proceedings of the 28th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '20), November 8– 13, 2020, Virtual Event, USA. ACM, New York, NY, USA, 4 pages. https://dl.acm.org/doi/10.1145/3368089.3417043

- **Md Sohel Ahmed**, Hiroshi Morita (2018) An Analysis of Housing Structures' Earthquake Vulnerability in Two Parts of Dhaka City. Sustainability (Impact factor: 1.78) 10:1106. https://www.mdpi.com/2071-1050/10/4/1106

- **Md Sohel Ahmed**, Hiroshi Morita (2019): An Analysis of Essential facilities and Occupancy class for earthquake preparedness in Two Parts of Dhaka City. Int. J. Earthquake and Impact Engineering 2(4) 339-359. https://www.inderscienceonline.com/doi/abs/10.1504/IJEIE.2018.099364

- **Md Sohel Ahmed**, Hiroshi Morita (2018) Housing structure analysis for earthquake disaster preparedness, 16th European conference on Earthquake, 18-21 June, 2018. Thessaloniki, Greece.

- **Md Sohel Ahmed**, Hiroshi Morita (2017) Earthquake disaster management analysis in Dhaka. In: Humanitarian Technology Conference (IHTC), 21-22 July, 2017. Toronto, Canada, pp 46–50.

PROFESSIONAL LINKS

GitHub Profile: https://github.com/sohel10
Publication Link: https://scholar.google.com/citations?view_op=list_works&hl=en&user=0RpVX20AAAAJ
LinkedIn Profile: https://www.linkedin.com/in/sohelcu06/

WORK ELIGIBILITY

US Permanent Resident (Green Card Holder)