Hala Arar

705-977-6821 | hala.arar.02@gmail.com | linkedin.com/in/hala-arar | github.com/halaarar | halaarar.github.io/halawebsite

SUMMARY

Data Scientist with a background in biomedical science and hands-on experience building NLP and machine learning systems. Developed end-to-end pipelines using PubMedBERT and XGBoost, deployed cloud-based solutions on AWS, and built interactive dashboards for clinical and public-sector use cases. Passionate about translating complex scientific and operational data into clear, actionable insights.

SKILLS

- **Languages:** Python, R, SQL, Bash
- **ML & NLP:** scikit-learn, XGBoost, LightGBM, TensorFlow, PyTorch, spaCy, Hugging Face, transformer models
- **Data Engineering:** ETL, data wrangling, REST APIs, SQL optimization
- **Tools:** Git, GitHub Actions, Jupyter, VS Code, Docker, AWS (S3, EC2, SageMaker)
- **Visualization:** Shiny, Plotly, Dash, geospatial dashboards

EXPERIENCE

Data Scientist – Capstone Project **Apr 2025 – Present**

University of Helsinki, Finland (Remote)

- Built a two-stage PubMedBERT pipeline for automated detection and classification of protein autoregulatory mechanisms across biomedical literature, achieving 96.0 percent accuracy in Stage 1 and 96.2 percent macro-F1 in Stage 2.
- Deployed the model across 252,880 unseen PubMed abstracts to generate the largest resource to date for predicted autoregulatory mechanisms and built an interactive Shiny app supporting search, filtering, and ontology-linked exploration.

Data Scientist – Research Collaboration **Jul 2025 – Oct 2025**

University of British Columbia, Vancouver, BC (Remote)

- Developed XGBoost and adversarial LLM models to classify clinical gene variants using structured patient features, improving consistency between ACMG categories and reducing borderline classifications.
- Engineered domain-specific features aligned with cardiogenetics guidelines and collaborated with geneticists to design interpretable ML workflows for future clinical decision support.

EDUCATION

Master of Data Science **Sep 2024 – Jun 2025**

University of British Columbia, Vancouver, BC

- Core Coursework: Advanced ML, Causal Inference, Data Viz, Cloud & Database Systems

Bachelor of Science, Biomedical Sciences **Sep 2020 – Apr 2024**

Trent University, Peterborough, ON

- Minor in Mathematics | Renewable Entrance Scholarship | Dean's Honour Roll

TECHNICAL PROJECTS

SOORENA: Autoregulatory Mechanism Detection (GitHub)

- Built a two-stage PubMedBERT NLP system achieving 96.0 percent accuracy in detecting autoregulation and 96.2 percent macro-F1 across seven biological mechanism classes
- Ran inference on 252,880 PubMed abstracts and developed an interactive Shiny dashboard with ontology-linked search and confidence scores

Bank Marketing Predictions (GitHub)

- Built a logistic regression model using 45,000+ customer records to predict term deposit subscriptions, improving targeting precision by 20 percent
- Delivered insights that reduced outreach waste and improved campaign ROI

NYPD Arrest Tracker App (GitHub)

- Developed a Plotly- and Pandas-based dashboard to visualize arrest patterns by borough, precinct, and offense category
- Supported NYC government analysis with dynamic, filterable geospatial visualizations

PUBLICATIONS

Arar, H., He, Z., Zhou, Y., Zhang, M., Nickchi, P., & Jafari, M. (2025). *SOORENA: A Two-Stage Deep Learning Pipeline for Automated Extraction of Autoregulatory Mechanisms in Biomedical Literature.* bioRxiv **(preprint).**