905-462-8303 | LinkedIn | GitHub | **HERAA MUQRI** | DevPost | heraa.muqri@mail.utoronto.ca | Website

EDUCATION

Bachelor Of Applied Science, Engineering Science + PEY Co-op | Machine Intelligence Major | University of Toronto | Expected June 2027

Relevant Coursework: Data Structures & Algorithms, Probability & Statistics, Machine Learning, Neural Networks, Calculus, Linear Algebra

Languages: Python, R, SQL (Postgres, MySQL), MATLAB
Libraries: PyTorch, Tensorflow, Keras, Scikit-learn, Pandas, Numpy, Seaborn, Matplotlib, Sklearn, TabPFN
Tools: Docker, Streamlit, PowerBI, Tableau, Jupyter Notebook, Jira, Git

WORK EXPERIENCE

Data Engineer Intern — Aug 2025 - Current
Xylem, *Toronto, Canada*

- Developed an end-to-end generative AI web tool (Gemini API, Python, LangChain, Streamlit, and Tesseract OCR), designed few-shot prompts, and tuned model parameters (e.g., temperature, top-p) to improve accuracy to 88%, enabling automated extraction of instrument specifications from PDFs, drawings, and Excel files, driving process optimization and reducing manual review time for 300+ engineers by 60%.
- Integrated Gemini API with Excel MCP server and RESTful APIs to automate batch file processing and streamline data ingestion.
- Deployed the solution into production across multiple workflows in compliance with internal standards, increasing project efficiency.

Operational Strategy and People Analytics Extern — Jun 2025 - Aug 2025
Extern, *Remote*

- Performed web scraping and sentiment analysis using Python natural language processing libraries like TextBlob, NLTK on 135+ Glassdoor reviews & 20 YouTube videos to identify operational challenges in fulfillment centers.
- Performed data analysis and visualization to create business briefs highlighting retention insights; results informed employee experience strategy.

Data Engineer Intern — May 2024 - Aug 2024
Xylem, *Toronto, Canada*

- Automated A/B testing on 200+ time-series datasets using Python and MATLAB signal processing toolkit, reducing analysis time by ~90%.
- Automated calculation of confusion matrix, ROC curves, and determined optimal data threshold using MATLAB script and Excel.
- Conducted QA/QC testing experiments on electromagnetic system performance of pipe inspection tools in a simulated environment.
- Collaborated with engineers to identify bottlenecks in testing and optimize the tool based on data, reducing signal noise by ~87%.
- Wrote detailed statistical reports on testing results, visualizing data through charts to help the engineering team's decision making process.

PUBLICATION

Towards a Reinforcement-Learning Based System for Adjusting Medication to Minimize Speech Disfluency — **Feb 2024**

- Collaborated with 17 students to collect and preprocess YouTube video data on speech disfluency and research medications to develop a reinforcement learning (RL) model to optimize medication adjustments for minimizing speech disfluency using RL techniques.

PROJECTS

Ensemble ML Model for Exoplanet Detection | GitHub | App | *Python, R, HTML, CSS, JavaScript, Flask, Three.js, Scikit-learn* — **Oct 2025**

- Preprocessed 9,500+ datapoints with 43 features using median imputation, one-hot encoding, and feature scaling for robust model training.
- Trained and tuned a Stacking Ensemble (Random Forest, Gradient Boosting) ML models using GridSearchCV and 5-fold cross validation, achieving 81.2% classification accuracy and conducted feature importance analysis (Linear Regression) to identify key predictors.
- Collaborated in a team of four to deploy the model via a Flask API with Three.js visualization to render real-time interactive 3D planet orbits.

ML System for Personalized Study Advice | GitHub | Research Paper | *Python, Scikit-learn, Pytorch, Tensorflow* — **Jan – Apr 2025**

- Collaborated in a team of three to develop an end-to-end personalized ML models system to recommend study habits to improve student's grades by training and tuning an end-to-end stacking ensemble (TabPFN + kNN) using GridSearchCV, achieving an R^2 of 0.953.
- Performed data analysis on 2,392 students' data using Python (Pandas, Seaborn, Matplotlib) and conducted feature importance analysis (Linear Regression) to identify key features influencing academic performance, guiding effective model training.
- Integrated machine learning model with a GPT-4 based chatbot GUI using Gradio and OpenAI API to create text-based study habit suggestions.

MLP Neural Network for Stress Detection Using EEG and ECG | Research Paper | GitHub | *Python, Tensorflow, Scikit-learn* — **Jul 2024**

- Trained and tested an MLP model and performed hyperparameter testing to tune model performance with a team of 2 to classify stress based on EEG and ECG signals using Python (Tensorflow, Sklearn, Numpy, Sci-kit, Pandas) and Jupyter Notebook.
- Prepared an abstract, poster presentation, and report presented at Microsoft Toronto office and published in STEMFellowship Journal.

Decision Tree Classifier for Undiagnosed Disease Treatment Prediction | Research Paper | GitHub | *Python, Scikit-learn* — **Jul 2023**

- Trained and tested a decision tree classifier using the data of undiagnosed disease patients from Harvard University's Disease Network achieving 88% accuracy by collaborating with a team of 3 and using Python (Numpy, Pandas, Scikit-learn, Sklearn)
- Prepared a report and abstract on results published in the STEMFellowship Journal.