# ALBERTA ARABA JOHNSON

**Email: arabaa.johnson@gmail.com**
**LinkedIn | GitHub**

## PERSONAL SUMMARY

I am a Ph.D. student in Statistics with a strong interest in improving health research and patient care through better data analysis. My current interest and work focus on developing scalable Bayesian Double Machine Learning (BDML) methods to estimate treatment effects from high-dimensional healthcare data. I am especially interested in tackling the challenges that come with real-world health data, such as confounding, noise, and complexity, and building methods that help us draw more reliable and interpretable conclusions. By combining ideas from causal inference, Bayesian modeling, and machine learning, I aim to support more accurate, data-driven decisions in clinical and public health settings.

## EDUCATION

- **Ph.D. in Applied Statistics**, *University of Northern Colorado* — Jan 2025 – Present
- **M.Sc. in Applied Statistics and Data Science**, *University of Texas Rio Grande Valley* — Aug 2022 – May 2024
- **B.Sc. in Actuarial Science**, *University of Cape Coast* — Aug 2017 – Oct 2021

## RESEARCH INTEREST

Causal Inference, Survival Analysis, Clinical Trials, Infectious Disease Modeling, Bayesian Statistics, Generalized and Mixed Effect Models, Machine Learning

## TECHNICAL SKILLS

**Languages & Software:** SAS, R, Python, SQL, SPSS, LaTeX, Microsoft Office

## WORKING EXPERIENCE

- **Research & Statistical Consultant** — Jan 2025 – Present
  *Research Consulting Lab, University of Northern Colorado*
  - Provide consultation to researchers in more than ten disciplines, including biology, nursing, and psychology, helping with data management, research design, data analysis, and interpretation to optimize statistical analysis.
  - Conduct hands-on training in R, SAS, and SPSS, and guide researchers in statistical techniques that contribute to projects and dissertations.

- **Statistics Tutor** — Jan 2025 – Present
  *Statistics Support Center, University of Northern Colorado*
  - Provide tutoring services on a variety of statistics courses to both undergraduate and graduate students.

- **Graduate Research Assistant** — May 2023 – May 2024
  *School of Mathematical and Statistical Sciences, The University of Texas Rio Grande Valley*
  - Synthesized domain literature.
  - Developed and implemented Classical Bayes, Maximum Likelihood, and Empirical Bayes estimators for the epidemic reproduction number ($R_0$), including simulation studies, performance assessment via regret risk, and estimator monotonicity using Van Houwelingen and Isotonic Regression methods.
  - Prepared technical reports and presented findings via posters and seminars.

- **Graduate Teaching Assistant** — Aug 2022 – May 2023
  *School of Mathematical and Statistical Sciences, The University of Texas Rio Grande Valley*
  - Provided technical support to undergraduates in MATH 3343: Introduction to Mathematical Software using Matlab, C++, and Python.
  - As a course instructor for MATH 1342: Elementary Statistical Methods, I created a learning environment that encouraged student involvement and participation.
  - Held regular office hours for academic guidance and clarification.

- **Teaching and Research Assistant** — Oct 2021 – Aug 2022
  *Department of Statistics, University of Cape Coast, Ghana*
  - Organized tutorials for students in Further Probability, Life Contingencies, and Survival Analysis.
  - Aided students in their undergraduate theses and projects.
  - Assisted in research-related tasks, particularly in the area of Statistical Data Analysis.
  - Performed administrative activities such as taking attendance, grading assignments, and maintaining student records.

- **Data Science Intern** Jun 2021 – Aug 2021
  *Amalitech Service GMBH, Ghana*
  - Used Python libraries BeautifulSoup and Scrapy to extract data on the German IT job market.
  - Conducted market research on Ghana's SME and IT job landscape.
  - Performed statistical analysis and visualization of the German IT job market to compare with IT jobs in Ghana.
  - Presented data-driven reports to enhance strategic decisions.

- **Data Management and Analysis Intern** Jun 2019 – Aug 2019
  *Prime Insurance, Ghana*
  - Developed an Excel-based customer data model for efficient file retrieval.
  - Generated and analyzed weekly Excel reports to improve operational efficiency.

## PUBLICATIONS AND PROJECTS

- Eshun, S., **Johnson, A. A.**, & Essieku, R. (2023). *COVID-19 Tweets Sentiment Classification: A Deep Learning Model with Stacking Ensemble for Improved Performance.* Accepted in the International Journal of Scientific Research in Mathematical and Statistical Sciences

- **Longitudinal Analysis of Asthma Treatment Using Mixed Models and GEE**
  - Conducted a longitudinal analysis of clinical trial data to assess the efficacy of asthma drugs. Applied and compared advanced statistical methods, including repeated-measures mixed models and Generalized Estimating Equations (GEE), to assess treatment, time, and interaction effects on FEV measurements. Validated model assumptions and ensured optimal model fit for robust inference in a clinical research context.
  - Compared covariance structures (UN, CS, AR(1), Toeplitz) using AIC, BIC, and QIC for optimal model fit. Validated model assumptions through residual diagnostics and ICC analysis. Identified the unstructured covariance matrix as the best for mixed models and AR(1) as the optimal for GEE.

- **Breast Cancer Diagnosis Using Machine Learning**
  - Evaluated nine supervised machine learning models including logistic regression, KNN, SVM, random forest, neural networks, and gradient boost on the Wisconsin Breast Cancer Dataset using R. Performed model tuning, 10-fold cross-validation, and feature selection using AIC/BIC criteria.
  - Compared model performance using sensitivity, specificity, F1 score, and balanced accuracy. KNN and neural networks outperformed others, achieving a balanced accuracy of 97.73% and an F1 score of 96.30%. Delivered reproducible results and visualizations using R and data science best practices.

- **Bayesian Estimation of Reproduction Numbers in Epidemics**
  - Developed Classical Bayes, Maximum Likelihood, and Empirical Bayes estimators under the square error loss function to estimate the basic reproduction number ($R_0$) from outbreak size distributions modeled using generalized Poisson and Galton-Watson branching processes.
  - Introduced monotonization techniques (Van Houwelingen and Isotonic Regression) to enhance estimator stability. Conducted simulation studies comparing estimator performance in terms of Bayes risk and regret risk, demonstrating that monotonized estimators achieved superior accuracy.

- **Ensemble Deep Learning Model for Sentiment Classification of COVID-19 Tweets**
  - Developed a meta-classifier deep learning framework for sentiment classification using the COVID dataset (90K tweets). Designed and trained LSTM, BiLSTM, and GRU models, and implemented an ensemble strategy combining BiLSTM outputs with a Random Forest meta-classifier. Achieved 99.97% accuracy, outperforming state-of-the-art baselines (e.g., DistilBERT at 96.66%).
  - Employed advanced preprocessing (tokenization, lemmatization, embedding), numerical transformation, and visualization techniques. Evaluated models using F1-score, precision, recall, and confusion matrices.

## HONORS AND AWARDS

- **Beatrice Heimerl Educational Research and Statistics Graduate Fellowship** Jun 2025
  *University of Northern Colorado*

- **ASRM Recognition – Graduate Research** Apr 2025
  *University of Northern Colorado*

- **College of Science Dean's Graduate Fellowship** Aug 2022 – May 2024
  *The University of Texas Rio Grande Valley*

- **MASEB Award for the Overall Best Female Student (Actuarial Science) in the Department of Statistics for the School of Physical Sciences** Dec 2021
  *University of Cape Coast*

## RELEVANT COURSEWORK

- **University of Northern Colorado**
  Statistical Methods, Advanced Research Methods, Experimental Design, Applied Time Series Analysis, Advanced Statistical Data Analysis, Applied Statistics
- **The University of Texas Rio Grande Valley**
  Biostatistics, Statistical Learning, Statistical Methods, Stochastic Processes, Neural Networks & Deep Learning, Time Series, Python for Statistics, Data Mining & Warehousing, Numerical Analysis
- **University of Cape Coast**
  Linear Models, Statistical Inference, Probability Distributions, Credibility Theory, Survival Models, Time Series Analysis, Multivariate Distributions, Linear Models, Statistics in Medicine

## REFERENCES


- Randy Larkins (PhD)
  Assistant Professor
  University of Northern Colorado
  randy.larkins@unco.edu
  +1 970 351 1676
- George Yanev (PhD)
  Professor
  University of Texas Rio Grande Valley
  george.yanev@utrgv.edu
  +1 956 665 3632
- Samuel Essamuah Assabil (PhD)
  Senior Lecturer
  University of Cape Coast
  samuel.assabil@ucc.edu.gh
  +233 24 042 9164