**Linh D. Shinguyen, M.S.**

[Lnguyen1021@gmail.com](mailto:Lnguyen1021@gmail.com) | [www.linkedin.com/in/shinguyenlinh](http://www.linkedin.com/in/shinguyenlinh)

**RELEVANT EXPERIENCE**

|  |  |
| --- | --- |
| **Emerging Leaders in Data Science Fellow: National Institute for Allergy and Infection Diseases**  9/2021 - Present  Oak Ridge Institute for Science and Education (ORISE), Rockville, MD   * Collaborated with various teams of researchers and program officers to address their data science needs. * Curated documentation for each project to identify data location, bottlenecks, and analysis plans. * Organized communication between researchers and policy drivers to advance secondary research and data science efforts around NIAD. | **Technology:**  R, Python, R Shiny |

**HIGHLIGHTED PROJECTS**

|  |  |
| --- | --- |
| 1. **Deep learning Lung and Lung Lesion Segmentation (1/2023 – Current)**   *Division of Intramural Research: Integrated Research Facility (IRF)*   * Building and evaluating 3 deep learning architectures (UNET, UNET-R, SWIN-UNETR) to conduct image segmentation on non-human primate lung CT scans. * Designing and assessing deep learning model performances to identify lung lesions caused by SARS-COV2 in CT image scans. * Creating and implementing a deep learning model benchmark performance pipeline for CT image segmentation on lung and lung lesions to be utilized for clinical researchers & radiologists. | **Technology:**  Python, PyTorch, ITK-Snap, monai, |
| 1. **Multimodal analysis of Multiple Sclerosis (9/2022 – 11/2022)**   *Division of Intramural Research (DIR): Laboratory of Clinical Immunology and Microbiology (LCIM)*   * Conducted an exploratory analysis to characterize individuals with multiple sclerosis (MS) utilizing multi-modal omics data collected from the cerebrospinal fluid. (i.e. patient demographics, immunophenotyping, bulk RNA sequencing, & somamer scan data). * Investigated differences between healthy controls and patients with MS by applying several dimension reduction and clustering techniques (e.g. uniform manifold approximation and projections (UMAPS), t-distributed stochastic neighbor embeddings (t-SNE), principle component analysis (PCA)) to identify multimodal discriminatory features. * Applied dimension reduction and clustering techniques (e.g. UMAPs, t-SNE, PCA) to identify multi-omic features to discriminate between MS subtypes. | **Technology:**  R |
| 1. **Generalized Data Exploration and Modeling Tool (4/2022 – 8/2022)**   *Division of Clinical Research (DCR): Bioinformatics Research Branch (BRB)*   * Created an interactive exploratory and modeling tool utilizing R-shiny to facilitate quick analysis of tabular data. * The tool allows users to interactively visualize variable correlations, data transformations, and predictive model performance in an easy to use app. | **Technology:**  R shiny |
| 1. **Prognostic modeling of post-acute sequelae of SARS-Cov-2 (PASC) (4/2022 – 08/2022)**   *Division of Clinical Research (DCR): Bioinformatics Research Branch (BRB)*   * Assisted in univariate analysis to characterize differences among individuals who developed post-acute sequalae of SARS-COV-2 and those who did not among persons infected with COVID-19. * Created random forest and elastic net models to explore predictive risk and diagnostic factors of PASC utilizing shapley values. | **Technology:**  R |
| 1. **Estimation of data generation from grants utilizing research performance progress reports (RPPRs): A Pilot Natural Language Processing (NLP) Project (11/2021 – 4/2022)** *Division of Allergy, Immunology, and Transplantation (DAIT)*  * Collected, wrangled, and cleaned 900 unstructured RPPR documents to create an NLP gold standard corpus for model building. * Extracted ‘data generated’ entities to curate a training dataset from 250 documents. * Created a named entity recognition (NER) model to classify documents which ‘generated data’. | **Technology:**  Clinical Language Annotation, Modeling, and Processing (CLAMP) Toolkit, Python |
| 1. **Improving Customer Email Conversion Rate (3/2021 – 5/2021)**   *Vanderbilt Data Science Institute industry partner*   * Prepared 43 GB of client data for modeling through data wrangling, cleaning, feature engineering, parallelization techniques on Vanderbilts high-performance computing cluster. * Segmented customers based on similar purchasing behaviors using K-means clustering. * Led a team of 4 first year graduate students to improve client email conversion rate utilizing machine learning models in Python. | **Technology:** Python, PySpark, MlLib, sklearn, |
| 1. **Customer Base Exploratory Data Analysis (1/2021 – 3/2021)**   *Vanderbilt Data Science Institute industry partner*   * Led a team of 4 first-year graduate students to conduct an exploratory data analysis project to provide a client with insights for their digital product performance utilizing customer data. * Created a data dashboard in Tableau for the client to identify national trends and monitor product performance over time. * ***Identified 10%*** businesses which were not staying current with payments thus allowing the client to contact these businesses to collect payments. | **Technology:**  Tableau, R, tidyverse, ggplot |
| 1. **Sub-Volcanic Discolored Water Pixel Classification (1/2021 – 5/2021)**   *Vanderbilt Department of Earth and Environmental Sciences*   * Utilized SENTINEL-II, satellite data, to create a random forest model in Google Earth Engine, to identify discolored water pixels from sub-volcanic eruptions. * Created additional satellite spectral bands feature as well as hyper parameter tuned the random forest model to improve ***classification accuracy from 73.2% to 84.1%***. | **Technology:**  Google Earth Engine, javascript |
| 1. **Validation of Cervical Pre-Cancer Billing Claims Data (8/2019 – 2/2021)**   *Vanderbilt University Medical Center*   * Collaborated with a PhD Candidate to clean 8,500 ICD-9 and ICD-10 billing records from 2008-2017 for Davidson County, TN, using Python. * Developed a random forest classification model (scikit-learn) to ***accurately identify 89%*** of cervical precancer events using billing records and validated the model results using biopsy data. * Published in JNCI Cancer Spectrum. <https://doi.org/10.1093/jncics/pkaa112>. | **Technology:**  Python, Pandas, Sklearn |
| 1. **Catastrophic Loss Prediction (1/2020 – 5/2020)**   *Vanderbilt Data Science Institute industry partner*   * Collaborated with classmates and an industry partner in Nashville to predict catastrophic losses in specific at-risk regions for sparse severe weather events between 2008-2018. * Provided the industry partner with insights for financial impacts of severe weather events through exploratory data analysis. | **Technology:**  Python, pandas, sklearn, seaborn |
| 1. **Latin American Survey Interview Adherence (8/2019 – 12/2019)**   *Vanderbilt Data Science Institute industry partner*   * Worked with Latin American stakeholders to identify abnormal survey behavior for 60 interviewers and ***identified 10%*** of interviewers who ***exhibited abnormal behaviors*** that were retrained to adhere to survey protocol. * Created a KNN clustering model in Python utilizing one year of survey, geospatial, and time series data to create a surveyor profile to help the client further understand surveyor’s behavior. | **Technology:**  R, ggplot, tidyverse, leaflet |

**EDUCATION TECHNICAL SKILLS**

|  |  |  |
| --- | --- | --- |
| **M.S. in Data Science**  8/2019 - 5/2021  Vanderbilt University, Nashville, TN  Capstone: Classification of Sub-Volcanic Discolored Water Pixels Using Satellite Data; Advisor: Tushar Mittal, Ph.D., Kristen Fauria, Ph.D. |  | R, tidyverse, ggplot, tidymodels  Python, sklearn, pandas, numpy, seaborn  Feature engineering  ML, Keras, Pytorch, sklearn  PySpark, TensorFlow  Data analysis & visualization  Data wrangling & cleaning  Tableau, Tableau Prep  SQL, MongoDB, data querying, ETL  Git, Github, Version Control  Endnote, Zotero, Reference management  LaTeX, Microsoft Word, Powerpoint, Excel,  Zoom, Cisco Webex, Microsoft Teams |
| **Certificate in Data Analytics and Visualization**  6/2018 - 11/2018  Georgia Institute of Technology, Atlanta, GA |  |
| **B.S. in Health Promotion and Behavior**  8/2013 - 5/2017  University of Georgia, Athens, GA |  |