**NGUYEN THAI BAO** – Intern Data Science

Phone: +84898489885 Email: baonguyenthai580@gmail.com Github: <https://github.com/baobao1911>

# EDUCATION

|  |  |
| --- | --- |
| **University of Science, Vietnam National University Ho Chi Minh City** | |
| Bachelor’s Program Specializing in High-Quality Data Science | 10/2020 – Present  (Expected graduation date) |
| GPA: 3.02   * Thesis : “Identyfy Prostate Cancer from Pathology Images using Pyramid Scene Parsing Network”   (Scored 9.3/10) | |

# TECHNICAL SKILLS

|  |  |  |  |
| --- | --- | --- | --- |
| * Machine Learning | * Python | * Pytorch, TensorFlow | * WebAPI |
| * Data Visualization | * PostgreSQL, MySQL | * Data Analysis | * Linux |
| * Data Warehouse | * ETL Processes | * Spark/ Hadoop | * Kafka |
| * MLflow | * Selenium /   BeautifulSoup | * Data Leak | * MinIo |

# PROFESSIONAL EXPERIENCE

|  |  |
| --- | --- |
| **Prostate Cancer Segmentation from pathology images** | July 2023 – December 2023 |
| * Programming language: Python | |
| * Frameworks: Pytorch, Albumentations, Jupyter Notebook, Numpy, Math, FlashAPI. | |
| * Github: [baobao1911/Prostate-Cancer-Segmentation-from-pathology-images](https://github.com/baobao1911/Prostate-Cancer-Segmentation-from-pathology-images) | |
| * Description : This project focuses on creating an AI model to automatically identify cancerous areas in prostate cancer images from the 2019 MICCAI Automatic Prostate Gleason Grading Challenge. The project will improve the PSPNet model’s performance through various enhancements and optimizations. The ultimate goal is to achieve more precise and dependable segmentation results, which will assist in diagnosing and planning treatment for prostate cancer. | |
| * Tasks: preprocessing data type H&E image , make balance, fine tuning pre-trained backbone ResNet,   modular improvements base on PSPNet , training and testing | |
|  | |
| **Features storage** | February 2024 – March 2024 |
| * Programming language: Python | |
| * Frameworks: Apache Airflow, PostgreSQL, Apache Spark, Apache kafka, Apache Flink. | |
| * Github: [baobao1911/Features\_storage](https://github.com/baobao1911/Features_storage) | |
| * Description: This project is dedicated to the development and deployment of a sophisticated feature store system, specifically designed to manage and serve trip records from the Yellow Taxi fleets of NYC. The system is engineered to cater to the unique requirements of each data stream, ensuring optimal performance and efficiency. Leveraging cutting-edge technologies and services such as PySpark, PostgreSQL, Flink, Kafka, DBT, and Airflow, the project aims to deliver a robust and reliable solution. | |
|  | |
| **Web scraping and Visualize** | February 2023 – April 2023 |
| * Programming language: Python | |
| * Frameworks: Selenium, Pandas, Matplotlib, Jupyter Notebook. | |
| * Github: [baobao1911/Web-Scraping-and-Analysts](https://github.com/baobao1911/Web-Scraping-and-Analysts) | |
| * Team size: 3 | |
| * Description: In this project, use the Selenium library to crawl data from a web page (WhoScored.com). After that, I processed the data to bring it back to a suitable format for visualization using the Matplotlib library. The result is a graph and valuable comments about the crawled data. | |
| * Tasks: Crawl data from website using selenium and analyze two specific aspects | |

# CERTIFICATIONS

|  |  |
| --- | --- |
| * Graduation course Data Visualization University of Illinois at Urbana-Champaign [coursera.org/share](https://coursera.org/share/3e418464ca893b461751ef5dcd1cb445) | April 2023 |
| * Graduation course Text Retrieval and Search Engines University of Illinois at UrbanaChampaign [coursera.org/share](https://coursera.org/share/71dbd6a1df214031eb90ce5852b94d58) | May 2023 |