

Wei-Han (Zoe) Hsu

• +1-332-249-9798 • zoe70416@gmail.com • [linkedin.com/in/weihanz-hsu/](https://www.linkedin.com/in/weihanz-hsu/) • zoe70416.github.io

Education

New York University Master of Science in Data Science (GPA: 3.7/4.0) Relevant Coursework: Machine Learning, Big Data, Deep Learning, Natural Language Processing, Computer Vision, Cloud Computing	New York, NY 2022-2024
Taipei Medical University B.S. in Medical Laboratory Science and Biotechnology (GPA: 3.9/4.0)	Taipei, Taiwan 2016-2020

Experience

Data Scientist New York University IT	New York, NY Feb 2023 - May 2024
<ul style="list-style-type: none">Spearheaded development of a TensorFlow Autoencoder-based anomaly detection system for the NYU research network, achieving a 73% F1 score in identifying network anomalies.Conducted data mining, data analysis (EDA), and feature engineering on a 100GB dataset using Pyspark/HPC.Engineered ETL scripts for ingesting unstructured network data, enhancing data pipeline efficiencyImplemented MLOps best practices for deployment using MLflow, Docker, and Kubernetes.	
Commercial Analytics/ Data Scientist Intern Medidata Solutions, Inc.	New York, NY May 2023- Aug 2023
<ul style="list-style-type: none">Engineered a Tree-based Marketing Mix Model (Random Forest) to optimize campaign strategies (including phone, in-person, and email outreach) projecting a 30% increase in sales to target customers.Implemented A/B testing methodologies to evaluate and enhance sales team performance, collaborating cross-functionally to extract comprehensive insights.Utilized Snowflake, PySpark, and AWS for data analysis and model development, creating Tableau dashboards to communicate complex information to stakeholders.	
Data Science Research Assistant Sleep Center of Shuang-Ho Hospital	Taipei, Taiwan Aug 2020- Mar 2021
<ul style="list-style-type: none">Engineered a multi-classification model for sleep apnea subtypes, reducing clinical decision-making time from 2-3 months to 3 weeks.Achieved 80%+ accuracy using ML algorithms (Random Forest, LR, SVM, etc.) with scikit-learn.	
Medical Technician/ Data Analyst aetherAI	Taipei, Taiwan Jul 2018- Jun 2022
<ul style="list-style-type: none">Spearheaded data quality initiatives, resulting in a 12% accuracy improvement for computer vision models used in image classification tasks.Used SQL to enhance the quality of labeled image datasets for machine learning.	

Selected Projects

Multimodal Report Generation PyTorch, BigQuery, LLM, Huggingface, CV, NLP, ViLT, HPC	[GitHub]
<ul style="list-style-type: none">Developed a multimodal system for automatic report generation from images, achieving a BLEU score of 0.34 using BLIP2 and prompt engineering, enhancing workflow efficiency.Fine-tuned large language models (Vicuna) on HPC using data parallel computing strategies in PyTorch.Preprocessed large amounts (567GB) of images and text from electronic health records.	
Credit Card Transaction Fraud Detection Python, sklearn, AWS (S3, Sagemaker)	[GitHub]
<ul style="list-style-type: none">Developed an ML model with 0.98 ROC score for fraud detection, leveraging AWS S3 and SageMaker for data storage, model training, and deployment.	

Technical Skills

Programming	Python, SQL, R, Bash
Libraries	Spark, Pyspark, Pytorch, Tensorflow, Pandas, Numpy, Matplotlib, sci-kit learn, Langchain, Hugging Face Transformer
Tools	AWS (S3, EMR, EC2, Sagemaker), HPC, Hadoop, Docker, Kubernetes. Tableau, Git,