KEVIN YEN

Phone: 716-750-8621 ◇ Email: kevinyen@buffalo.edu
Homepage: https://jaihuayen.github.io/

SUMMARY

Biostatistics Ph.D. Candidate and former Senior AI Software Engineer with over 4 years of experience in observational studies and training machine learning models. Proficient in advanced R and Python programming, with a strong background in large-scale data analytics and machine learning model deployment. Passionate about bridging advanced functional and longitudinal data analysis with innovative machine learning methodologies to drive measurable impact.

EDUCATION

The State University of New York at Buffalo *August 2024 - May 2028 (expected)*
Ph.D Candidate in Biostatistics

National Taiwan University *August 2017 - August 2019*
Master of Science in Agronomy (Biometry Division)

National Tsing Hua University *August 2013 - June 2017*
Bachelor of Science in Quantitative Finance

EXPERIENCE & RESEARCH

The State University of New York at Buffalo *August 2025 - Now*
Graduate Student Researcher: Semi-Parametric Non-Linear Mixed-Effects Random Forest (SP-MERF) on Clustered Data

- Designed the SP-MERF framework, unifying random forests and factor-smoothing splines in an iterative backfitting loop to capture non-parametric global fixed effects and non-linear cluster-level deviations without manual interaction specification
- Evaluated predictive performance on the Mayo Clinic Primary Biliary Cirrhosis (PBC) cohort (1,833 records; 282 patients) via 5-fold cross-validation, outperforming linear mix-effects model and tree-based baselines by up to 31.9% in prediction error
- Executed large-scale simulation studies and released an open-source R package for SP-MERF on clustered and longitudinal data

Graduate Student Researcher: Custom-Ensemble Random Survival Forest in CRISP Cohort

- Engineered a iterative tree-level optimization architecture across 100 bootstrap-sampled survival trees in Python
- Evaluated model on the CRISP kidney disease cohort, outperforming standard forest-level tuned Random Survival Forests by 1.75% in C-index

QNAP *February 2022 - June 2024*
Senior AI Software Engineer

- Developed and optimized a 250-class ResNet50 classification model (70% F1-Score) and vision transformer-based OCR models (90% accuracy) to scale high-dimensional image search capabilities
- Integrated multimodal semantic search into local file engines using CLIP (Contrastive Language-Image Pre-Training), transitioning systems from rigid keyword matching to natural language query processing
- Engineered edge-device deployment workflows utilizing TPU/NPU acceleration, optimizing inference speed and data privacy under strict hardware constraints
- Mentored junior personnel on programming, project organization, and building OCR training dataset

TutorABC *May 2020 - February 2022*
AI Algorithm Engineer

- Reduced customer churn by 5% by designing a random forest ensemble model with 81% predictive accuracy and applying explainable AI (SHAP) to guide targeted business retention strategies
- Extracted key customer feedback using computer vision (YOLO, OCR) and natural language processing (TFIDF, BERT) models, enhancing data analysis for additional product features and improved user experience
- Accelerated data pipelines and ETL throughput by 50% by designing automated, end-to-end MLOps workflows using Apache Airflow

National Taiwan University *August 2017 - August 2019*
Research Assistant: Species Richness Estimation under Species Identity Error

- Conducted data cleaning on weed and plant cover species surveys in Soft Bridge County, Taiwan, and Grand St. Bernard Pass, Switzerland
- Derived an adjusted Chao2 statistical estimator to account for measurement error, successfully reducing underestimation by 15% and correcting a 10% overestimation bias in sparse data settings
- Designed and deployed an interactive R-Shiny web application to translate complex statistical inferences and multi-site survey data for non-technical stakeholders

SKILLS

- **Core Competencies:** Machine Learning (Random Forest, Generalized Additive Model, Explainable Methods), Deep Learning (Computer Vision, Natural Language Processing, Multimodal), Statistical Analysis (Causal Inference, Longitudinal Data Analysis, Functional Data Analysis, Survival Analysis, Clinical Trial, Missing Value, Structural Equation Modeling), Machine Learning Operations (MLOps), Edge Computing
- **Programming Languages:** Python, R, SQL (MySQL, MSSQL, NoSQL), SAS, C++, MATLAB, Golang
- **Machine Learning Tools:** Pytorch, Tensorflow, PaddlePaddle
- **Tools & Platforms:** Linux, Docker, Git, LaTex, Airflow, Terminal CLI (Gemini, Antigravity, Claude), MongoDB

PUBLICATIONS

*represents the corresponding author

Yen J.-H., Chiu, C.-H.* (2020). Richness estimation with species identity error. *Proceedings 62nd International Statistical Institute World Statistics Congress*, **Volume 6**, 401-408.

Manuscripts in Preparation:

Yen K J.-H.*, Landsittel, D. (2026) Semi-Parametric Non-Linear Mixed-Effects Random Forest on Clustered Data.

TEACHING EXPERIENCE

The State University of New York at Buffalo *August 2024 - Now*
Graduate Assistant & Teaching Assistant

National Taiwan University *September 2017 - January 2019*
Teaching Assistant & Conference Lecturer