Haike Yu

✉ haike.yu@outlook.com | github.com/RemMyFav | ☎ 647-762-0703 | linkedin.com/in/haike-yu

Education

Georgia Institute of Technology Augest 2025 - TBD
M.Sc. in Computer Science (Machine Learning Specialization) GPA: 3.00 / 4.00

University of Toronto Sept 2020 - June 2024
Honours B.Sc. in Computer Science (Specialist), Mathematics (Major) GPA: 3.09 / 4.00

Advanced Courses: Scalable Computing (A-), Image Understanding (A-), Machine Learning Capstone Design for Vision (A+), Software Engineering (A+), Operating System(A), Machine Learning (A-), Artificial Intelligence (A-), Mobile Robotics (A+), Neural Networks and Deep Learning (A)

Projects

DeepPHQ — Hierarchical Text Modeling and Multi-Model Analysis for PHQ-8 *Nov 2025 – Dec 2025*
Technical Paper and Code available: Repo Link

- Led the full research direction and system design, proposing a hierarchical (word-, sentence-, and dialogue-level) framework to study how textual scope influences PHQ-8 depression prediction.
- Built the complete data pipeline from raw DAIC-WOZ clinical interview transcripts, including text cleaning, participant ID reconstruction, balanced sampling, and unified dataset generation across all granularity levels.
- Designed controlled, architecture-agnostic comparisons across LSTM, RNN, CNN-based text models, and a custom Transformer encoder to ensure fair evaluation of contextual modeling capacity.
- Analyzed how increasing contextual aggregation affects prediction behavior, showing that dialogue-level representations better capture long-range emotional patterns while also amplifying dataset-specific biases.

UKF vs EKF Analysis – Probabilistic State Estimation Modeling *Mar 2024 – Apr 2024*
Technical Paper and Code available: Repo Link

- Compared Unscented Kalman Filter (UKF) and Extended Kalman Filter (EKF) for nonlinear state estimation, focusing on trade-offs between accuracy, runtime efficiency, and numerical stability.
- Designed a simulation pipeline that outputs noisy measurements from nonlinear systems and logs both ground-truth and estimated states at one-second intervals, enabling accurate evaluation of filter precision and computational efficiency.
- Implemented UKF and EKF from scratch using modular Python class structures, optimizing matrix operations with NumPy and ensuring numerical stability in Kalman gain updates and covariance propagation.
- Evaluated on three nonlinear systems—Van der Pol oscillator, Lorenz attractor, and Duffing equation—revealing that UKF delivers more accurate estimations under chaotic dynamics, while EKF is more computationally efficient.

Handy-the-Mystic-Hand – Random Forest Hand Recognition OS Control *Oct 2023 – Dec 2023*
Presentation Report and Code available: Repo Link

- Designed a gesture-based interface that allows users to control OS-level operations (click, drag, alt-tab, volume control) using only hand signs, targeting accessibility and physical strain reduction.
- Built a semi-automated image database: extracted 21 hand landmarks via MediaPipe, embedded samples using ResNet50, applied cosine similarity to remove redundancy, and auto-clustered gestures with minimal manual labeling.
- Trained a Random Forest classifier on normalized landmark vectors to detect gesture classes, achieving 82% validation accuracy and integrated OpenCV for visualization and real-time testing.
- Connected gesture recognition with OS command execution to create a real-time, camera-driven interaction system capable of controlling desktop environments without mouse or keyboard input.

CinemaScopeAI – Multi-Task Movie Trailer Frame Classifier *Mar 2024 – Apr 2024*
Code available: Repo Link

- Investigated whether a single trailer frame could predict a film's budget category and genre, treating the task as a multi-label, multi-task image classification problem.

- Scraped IMDb Top 250 movies using Selenium and BeautifulSoup, extracted key trailer frames via OpenCV, and embedded binary labels into filenames for streamlined data handling.
- Built a dual-head CNN from scratch using TensorFlow: shared convolutional base + two output heads (sigmoid for genre, softmax for budget), with custom losses, early stopping, learning rate scheduler, and tf.data pipeline.
- Conducted transfer learning using VGG16 as a frozen feature extractor, attaching a custom dense head and training with binary cross-entropy on a 15-bit multi-label target.
- Compared both custom CNN and VGG16-based methods, achieving up to 76% accuracy in joint genre and budget classification.

Professional Experience

Vosyn Inc. – Machine Learning Intern *May 2025 – Dec 2025*

Toronto, Canada

- Successfully integrated DiarizationLM based on paper into the company's GCP-based FasterWhisper + NeMo pipeline to support experimentation on improving Word Error Rate (WER) and Word Diarization Error Rate (WDER) using large language model (LLM)-enhanced speaker diarization.
- Independently built an experimental platform by sourcing raw audio data, performing data cleaning and formatting, and constructing a reference-based evaluation pipeline for automated WDER scoring.
- Proactively proposed a strategic shift in debugging direction: instead of further optimizing diarization models, isolated the root causes of high WDER in the Automatic Speech Recognition (ASR) module and initiated modular refactoring of the transcription stage using FasterWhisper.
- Tuned model parameters under tight GPU resource constraints, balancing inference speed, GPU memory, and transcription accuracy to ensure pipeline stability in production.
- Diagnosed and resolved instability issues in timestamp alignment caused by 'timestampcfcforcealigner', leading to a 25% improvement in WDER stability after successfully combining WhisperX and NeMo diarization outputs.

People's Bank of China – Data Analyst Intern *Jul 2023 – Aug 2023*

Guiyang, China

- Reclassified loan categories by querying and analyzing Oracle databases, improving high-risk loan identification accuracy by 13%, enhancing financial risk assessment.
- Developed internal REST APIs to streamline data access across analytical teams, improving workflow efficiency and risk model integration.
- Conducted exploratory graph modeling with Neo4j on financial transaction data to improve downstream ML-readiness and relational pattern mining.

Bank of Guiyang – Data Processing Intern *May 2023 – Jun 2023*

Guiyang, China

- Built scalable web scrapers using Scrapy to extract and store 100,000+ job postings for labor market analysis.
- Constructed a Python-based data pipeline to clean and deduplicate datasets, identifying 1,000+ relevant records.
- Deployed structured data using Flask on Alibaba Cloud, enabling real-time access for internal analytics teams.

Skills

Programming Language: Python, Java, TypeScript, C, JavaScript

Machine Learning: Logistic, Linear Regression, K-Means, Decision Trees, Random Forest, Clustering

Deep Learning: CNN, RNN, LSTM, Transfer Learning, ResNet, VGG, Large Language Models (LLMs)

Database: Redis, Cassandra, Neo4j, SQL, NoSQL, MongoDB

ML Tools: PyTorch, OpenCV, MediaPipe, Scikit-learn, Numpy, Matplotlib, Scrapy, TensorFlow

Cloud Services: AWS, Git, AliBaba Cloud, GCP

Other Tools: Bash/Shell Scripting, Docker, Latex