

# Zhangyue Shi

Tel: +1 (405) 762-2899

Email: [zhangyueshi97@gmail.com](mailto:zhangyueshi97@gmail.com)

Web: <https://shizhangyue.github.io/>

## CURRENT POSITION

### Bayer R&D Services LLC

Jul. 2023 – present

#### Data Scientist

- Utilize different machine learning models to predict different phenotypic traits using genotypic data using PyTorch, enhancing prediction performance compared to traditional linear model and providing insight for product development team.
- Deploy a novel tail-preserving loss function to make more accurate predictions at the tail region in AWS, improving performance in RMSE compared with mean square error loss and enabling high-performing lines discovery.
- Participate in code development and maintenance for corresponding python library in GitHub.

## EDUCATION

### Oklahoma State University (OSU)

Jul. 2023

Ph.D. in Industrial Engineering & Management, Minor in Statistics. GPA: 4.0/4.0

Stillwater, OK, USA

- *Dissertation Topic: Advanced Data Analytics-based Quality Assurance for Smart Manufacturing*

### Xi'an Jiaotong University (XJTU)

Jun. 2019

B.S. in Mechanical Engineering. GPA: 3.7/4.0

Xi'an, Shaanxi, China

## SKILL

**Programming:** Python, R, MATLAB, SQL

**DS Library:** Scikit-learn, PyTorch, TensorFlow, Gurobi, Pandas, Numpy

**Software:** Git, AWS, Google Cloud, Domino

**Hardware:** Raspberry Pi

**Quantitative Methodologies:** Machine Learning, Deep Learning, Bioinformatics, Genomic Selection, Statistical Inference, Knowledge Distillation, Continual Learning, Data Augmentation, Linear Programming, Time Series Analysis, Stochastic Process.

## PREVIOUS EXPERIENCE (SELECTED 5 OUT OF 9)

### Machine Learning Experience in Bioinformatics

- 1. Breeding Value Prediction via Genotype Data**, *Data Scientist Co-op*, Bayer Crop Science Jan. 2023 – Jun. 2023
  - Applied supervised machine learning methods (boosting, random forest, and neural network) in TensorFlow and scikit-learn to predict yield of maize using genotype data together with environment features (in Google Cloud, AWS, and Domino).
  - Identified feature importance among different genotype, soil, and weather features via saliency map and guided backpropagation.
- 2. Identification of Characteristics of Small Non-coding RNA with Aging**, *Research Assistant*, OSU Oct. 2021 – Jan. 2023
  - Found the correlation between small non-coding RNA with human aging via plasma and serum of healthy donors using machine learning methods, which could achieve 0.94 in  $R^2$  and  $RMSE \leq 3.7$  in test set based on the established age prediction model.

### Machine Learning Experience in Anomaly Detection

- 1. ML Model Performance Enhancement in Distributed System**, *Research Assistant*, OSU Nov. 2021 – Dec. 2022
  - Enhanced monitoring performance of neural networks in distributed manufacturing system via knowledge distillation (in PyTorch), which made performance of convolutional neural network (CNN) at data-poor agent have a 0.05 increment to 0.8 in f-score.
  - Distilled useful knowledge from CNN at data-rich agent to improve CNN at data-poor agents while preserving data privacy.
  - Collected time-series printing data for normal and anomaly condition at different agents and applied CNN to detect process anomaly.
- 2. Online Process Monitoring in Additive Manufacturing**, *Research Assistant*, OSU Feb. 2020 – Nov. 2021
  - Detected process anomaly during manufacturing via neural network (in TensorFlow), machine learning (in scikit-learn), and statistical method (in R), which could achieve 0.94 f-score in supervised monitoring and 85% true detection rate in unsupervised monitoring.
  - Extracted low dimensional features from high-dimensional time-series data using neural network (LSTM-autoencoder).
  - Applied supervised adaptive boosting monitoring; unsupervised one-class SVM and EWMA control chart monitoring from features.
- 3. Data Augmentation for Time-series Data in Manufacturing**, *Research Assistant*, OSU Aug. 2019 – Oct. 2021
  - Developed a data augmentation method based on generative adversarial neural network (in TensorFlow) to generate high-quality data; improved classification performance of logistic regression by 0.01 to 0.88 in f-score compared with model without data augmentation.
  - Proposed a filter layer in the generator to select augmented sample of high-quality based on time-regularized Hausdorff distance, which could discover underlying temporal relationship among the time-series data collected during manufacturing.

## SELECTED PUBLICATIONS (SELECTED 1 OUT OF 11)

- **Shi, Z.**, Mamun, A. A., Kan, C., Tian, W., & Liu, C. (2022). *An LSTM-autoencoder based online side channel monitoring approach for cyber-physical attack detection in additive manufacturing*. **Journal of Intelligent Manufacturing**.
- 11 published and accepted in total, 8 first-author papers. A full list of publications can be found [here](#).

## SELECTED HONORS AND AWARDS

- **Robberson Research and Creative Activity Grant (2 recipients per year)**, The Graduate College, Oklahoma State University, 2022.
- **Outstanding Graduate Student**, The College of Engineering, Architecture and Technology, Oklahoma State University, 2022.
- **National Science Foundation (NSF) Student Support Award**, North American Manufacturing Research Conference 49, 2021.
- **Data Challenge Competition Finalist**, Quality, Statistics and Reliability (QSR) Section, INFORMS Annual Meeting, 2019.
- **First Prize, National College Students Mathematical Modeling Competition**, Undergraduate Group Shaanxi Division, China, 2017.

## COMMUNITY ENGAGEMENT

**President**, INFORMS Student Chapter at OSU / **Reviewer**, Journal of Intelligent Manufacturing, Healthcare Analytics / **Member**, IISE, INFORMS