RESEARCH STATEMENT

Junxian Huang (junxian.huang@gmail.com)

My research interests span the areas of computer vision, bioinformatics and automated AI systems. Throughout my career in academia, industry and entrepreneurship, the top driving goal is to bridge the last-mile gap between theoretical research and real-world applications. A common thread in my research is in understanding and designing new machine learning techniques and building their interdisciplinary applications, either in the form of online web tools, or automated AI systems/machines. Specifically, my research focuses on solving challenging problems that arise from traditional fields, such as agriculture and manufacturing industry, in which, human labor is intensive yet inefficient and inaccurate, with human health and animal welfare being jeopardized in many circumstances, offering huge potentials for AI applications. Broadly speaking, my research belongs to the area of *Machine Learning*, which deals with the fundamental principles of data mining and modeling, and developing innovative, elegant algorithms for various challenging scenarios.

In my research methodology, I aim to identify key bottleneck problems, that can be solved in a more efficient, accurate, cost and environment friendly way, with the help of new machine learning techniques. Then I gather public data and collect additional domain data, sometimes with the software and hardware tools built by ourselves, and formulate new learning problems with various requirements coming from production. I further develop efficient algorithms to solve these key problems and validate the intellectual outcome in the real production settings.

Research in Agriculture and Manufacturing Industry

During the entrepreneurial gap of my academic career, I have personally visited many manufacturing factories, agricultural farms and companies, *etc.* The current level of automation and usage of AI technology is still very limited in these traditional fields. For example, when I visited one of the world's largest layer hatcheries in China, over 20 young female workers were standing on both sides of an endlessly moving line of chicks, differentiating their gender and health conditions with bare hands. Besides these production scenarios, even for researchers in these fields, much of the research work is done in a very inefficient way. Many problems in these areas share some similar challenges: 1) existing high-quality data is scarce, due to the lack of social and research attentions; 2) data collection is challenging, due to the complex nature of the problems and the lack of infrastructure support; and 3) data labeling is prone to errors, due to the lack of standardization and the difficulty in accurate measurement in real production environments. My previous research in agriculture and manufacturing industry has tackled with these challenges, and demonstrated both the richness of research questions arising from this area, as well as the practical contributions for several important applications.

Systems for Enhancing Poultry Breeding, Farming and Production

Chicken Phenotype Measurement: By discussing with my collaborators from Poultry Institute, Chinese Academy of Agricultural Sciences, I identified that efficient and accurate phenotype measurement is vital for enhancing poultry breeding, as existing measurement is mostly carried out manually or by expensive and complex equipments, which are often only suitable for lab usage. We designed and developed a simple yet effective machine learning based system to assist chicken phenotype field measurement. Before each measurement, researchers use a simple plastic stick to fix the distance between smartphone camera and a designed reference card, with 2cm $\times$ 2cm black cell for distance calibration and other color cells for color calibration. During the measurement, researchers use an Android app to take a photo using the same plastic stick to control camera-target distance and upload the photo to the cloud for analysis. We have developed efficient and accurate algorithms for measuring chick head organ parameters[1], feather parameters[2], pore parameters[3], egg parameters[6] and body parameters. The system is currently being used widely among researchers of the Poultry Institute and their external collaborators.
SmartEars – Diagnosing Respiratory Diseases Based on Chicken Vocalizations: We collaborated with one of the largest poultry companies in China, HuaYu Agricultural Science And Technology Co., LTD and after talking with their experts, I identified that detecting health problems of chicken flocks in a timely manner is essential for their production. Currently, their veterinarians need to walk a long distance every day inside the chicken farms and manually check chicken vocalizations during night hours, when the chicken flocks are in sleep and the background noise level is low, and identify abnormal sounds for early diagnosis of respiratory diseases. We

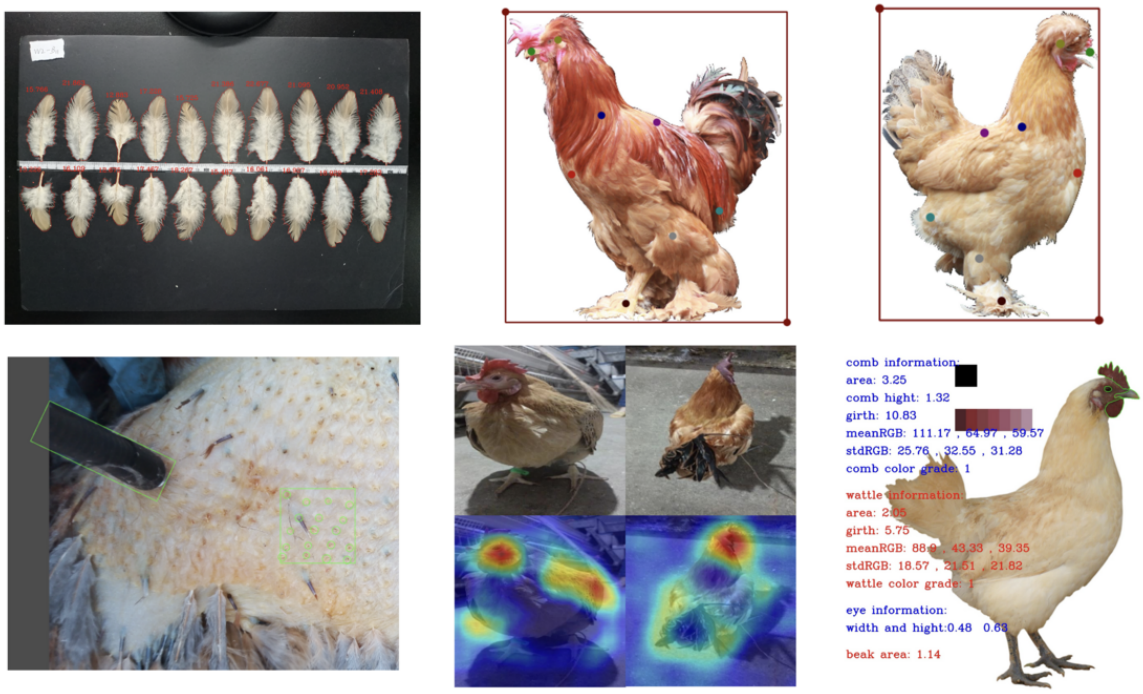


Figure 1: Chicken Phenotype Measurement

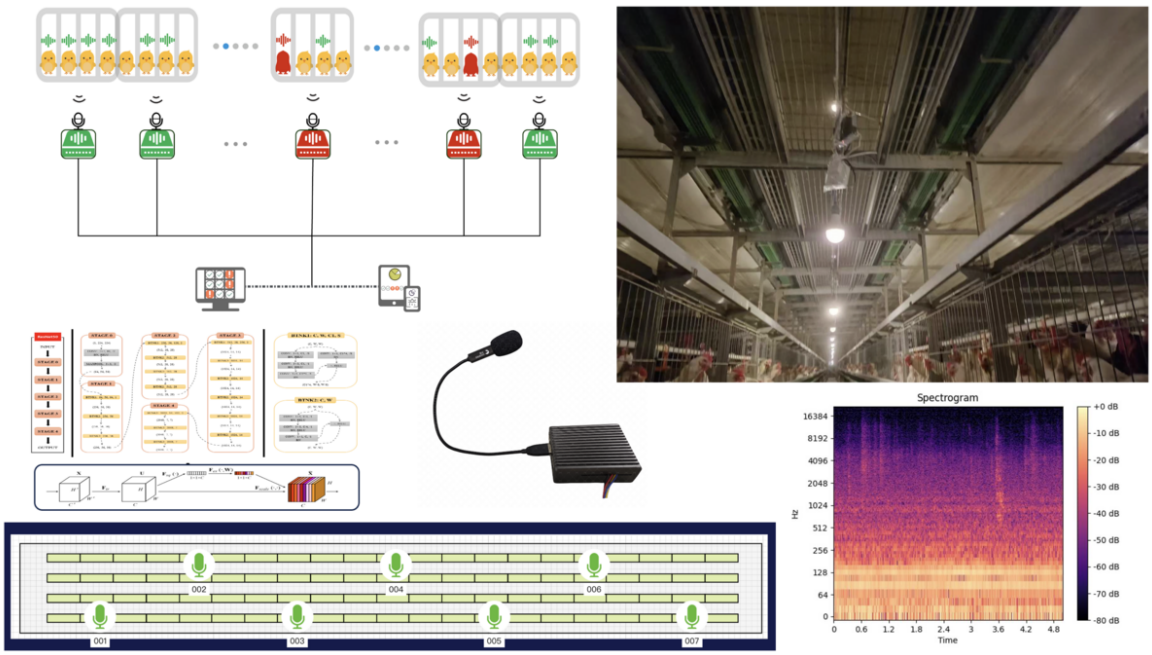


Figure 2: SmartEars – Diagnosing Respiratory Diseases Based on Chicken Vocalizations

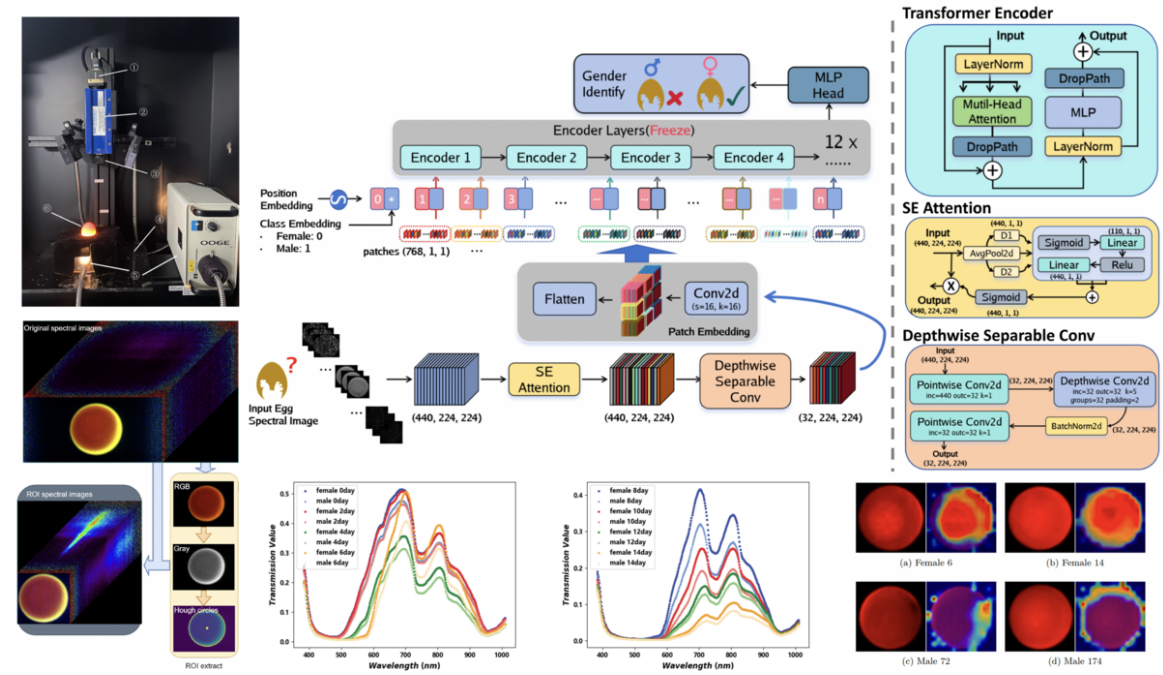


Figure 3: EggFormer – Nondestructive In-Ovo Sexing using Hyperspectral Imaging

developed SmartEars[7, 8], a system that collects audio samples periodically via edge computing devices placed in different locations of the chicken farms and equipped with an input microphone, and analyzes the audio samples to detect any disease symptoms. Veterinarians have helped us label over 40,000 audio samples, however, due to human error and the complexity of an audio sample in real production environments, the label quality is very low. We once carried out an experiment by asking three veterinarians to label the same audio data set, only around 40% of the labels reached consensus, because unlike many research datasets, the production audio consists of various environmental noises in additional to multiple chicken vocalization from difference distances mixing together. To deal with this challenge, we developed a new weakly supervised learning algorithm for audio classification based on Mel spectrograms, achieving 94% accuracy, and upon evaluation, our algorithm has demonstrated comparable, if not superior performance compared with domain experts. The system has been deployed in 3 chicken farms so far, and has analyzed over 3 million pieces of audio samples, and larger-scale deployment is expected within 2024.

EggFormer – Nondestructive In-Ovo Sexing using Hyperspectral Imaging: We explored the early gender identification of Hy-Line Sonia eggs using hyperspectral imaging and proposed EggFormer[12], a new deep neural network model based on Transformer. The culling of day-old male chicks post-incubation has raised animal welfare concerns worldwide and has also led to unnecessary costs during the incubation period. We observed that the widely used average spectrum representation might lose potential information when only one mean value is taken for each channel. Therefore, the motivation for this work is to propose a model that can extract more information from channel images and achieve better performance compared to existing methods. We investigated the spectral images at two-day intervals from day 0 to 14, focusing on day 10, which demonstrated the best accuracy of 95.4% using EggFormer with full-band images as input. Finally, we identified 22 most relevant wavelengths that retained the same performance by interpreting the internal mechanism of EggFormer, which is fewer than the 25 bands selected by CARS. This work provides a promisingly efficient and economical solution for hatcheries around the globe. We are working on building an automated prototype AI machine based on this technique with our collaborators.

Algorithms for Automated Quality Control in Manufacturing Industry

Together with my collaborators from Zhongke Kexin Optics and Electronics Co., LTD, we built automated AI machines for defect detection for various types of products, such as paper boxes, plastic bags, metal lids, *etc.* I was in charge of the design and development of all the software components, including camera control, AI

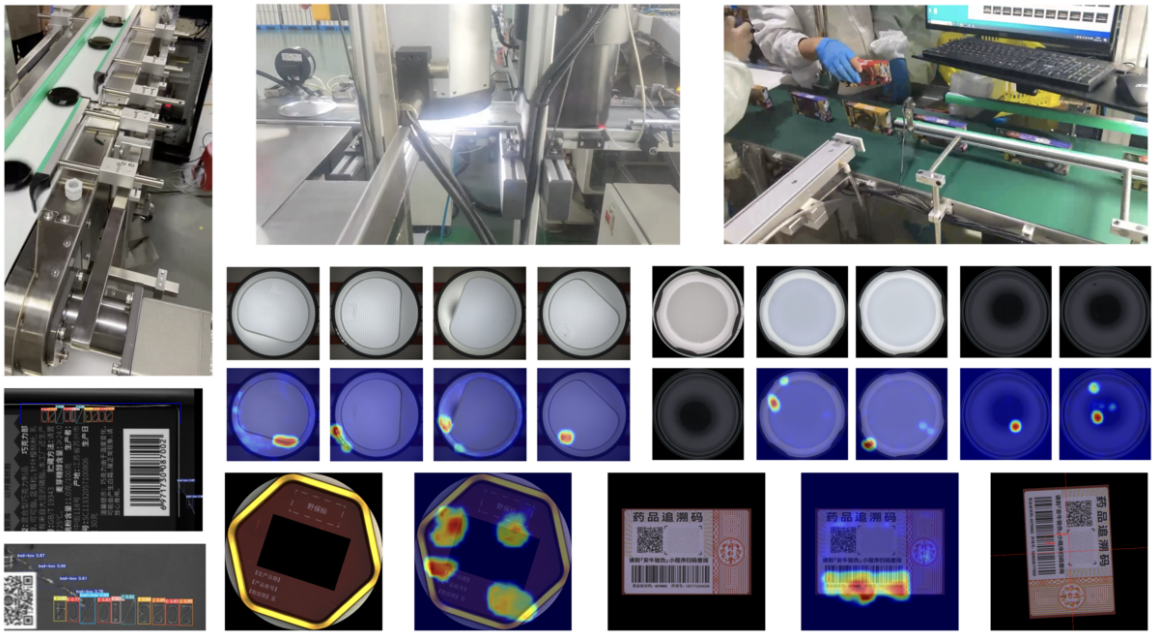


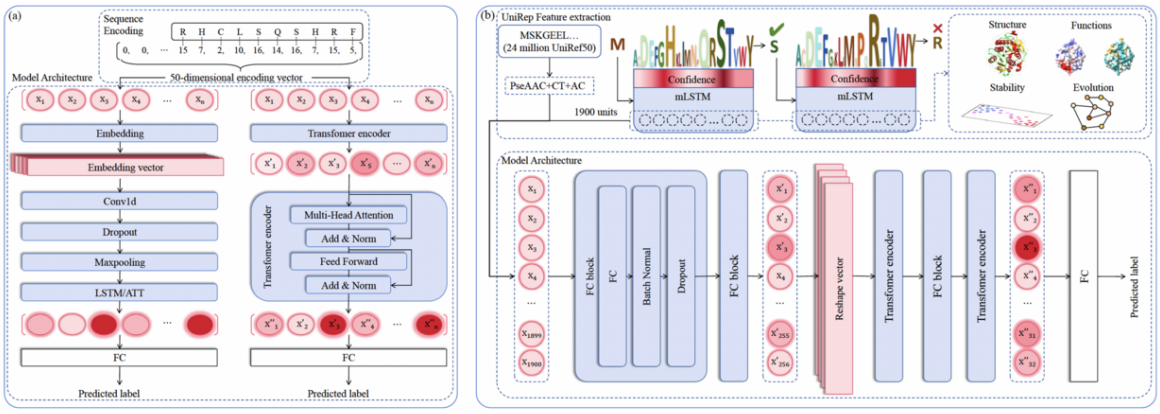
Figure 4: Applications for Automated Quality Control in Manufacturing Industry

detection, and network communication with different hardware components for eliminating defective products while transferring the remaining qualified ones. We designed a new machine learning algorithm based on deep neural networks, for efficient and accurate defect detection for over 20 products, with over 99% accuracy. The overall throughput of our machines can reach 180 products per minute. Our machines have been put into real production in a number of factories since the end of 2023, generating positive feedback. In the meanwhile, the deployment in the real production settings has enabled us to collect extensive data for future research, and given us a clearer picture of the key bottlenecks that prevent the manufacturing industry from upgrading in a faster pace.

Tools for DNA/Protein Analysis

By discussing with my collaborative researchers from College of Plant Protection, College of Animal Science, College of Life Sciences, College of Agriculture and College of Resources and Environmental Sciences in Nanjing Agricultural University, I realized that the tools for bioinformatics analysis equipped with the most recent advances in machine learning are in great demand. By summarizing their core requirements, we built some common tools to help them make analysis on DNA and protein based on the sequence and structural features, powered by specially tailored machine learning algorithms.

UniAMP – AMP Prediction with Inferred Information of Peptides: Antimicrobial peptides (AMPs) have been widely recognized as a promising solution to combat antimicrobial resistance of microorganisms due to the increasing abuse of antibiotics in medicine and agriculture around the globe. We proposed UniAMP[11], a systematic prediction framework for discovering AMPs. We observed that feature vectors used in various existing studies constructed from peptide information, such as sequence, composition, and structure, can be augmented and even replaced by information inferred by deep learning models. Specifically, we used a 1900-dimension feature vector inferred by an mLSTM model to demonstrate that such inferred information of peptides suffice for the task, with the help of our proposed deep neural network model composed of fully connected layers and transformer encoders for predicting the antibacterial activity of peptides. Evaluation results demonstrate superior performance of our proposed model on both balanced benchmark datasets and imbalanced test datasets compared with existing studies. Subsequently, we analyzed the relations among peptide sequences, manually extracted features, and automatically inferred information by deep learning models, leading to observations that the inferred information is more comprehensive and non-redundant for the task of predicting AMPs. Moreover, this approach alleviates the impact of the scarcity of positive data and demonstrates great potential in future research and



P. aeruginosa test dataset

Method	*TP*	*FP*	*TN*	*FN*	Pre	Rec	F_1-score	MCC
LSTM	826	41	99259	167	0.9527	0.8318	0.8882	0.8892
ATT	826	77	99223	167	0.9147	0.8318	0.8713	0.8711
BERT	848	94	99206	145	0.9002	**0.854**	0.8765	0.8756
LSTM&ATT&BERT	760	10	99290	233	**0.987**	0.7654	0.8622	0.8681
UniAMP (with PseAAC, CT and AC)	771	124	99176	222	0.8615	0.7764	0.8168	0.8161
UniAMP	848	61	99239	145	0.9329	**0.854**	**0.8917**	**0.8915**

Figure 5: UniAMP – AMP Prediction with Inferred Information of Peptides

applications. With the help of UniAMP, researchers from Poultry Institute has successfully discovered a brand new AMP from the metagenome of chicken faeces targeting *Salmonella* with a minimum inhibitory concentration (MIC) of 32 $\mu g/ml$.

RiceTFTarget – Rice Transcription Factor-Target Prediction Based on Coexpression: Transcription factors (TFs), also known as transacting factors, usually recognize the DNA cis-regulatory elements in the promoter regions of target genes to activate or repress expression. Identifying target genes of TFs or the TFs binding to target genes is crucial to address the biological functions and regulatory networks of these TF-target modules. However, the determining TF-target interaction is time-consuming and laborious. We constructed RiceTFtarget[10], an online web tool for robustly predicting TF-target pairs based on coexpression, pattern matching, and machine learning. Although some tools can be used for predicting cis-regulatory elements, RiceTFtarget is a tool for retrieving specific TF-target interactions in plants. We applied 46 TF binding site (TFBS) features to build 8 single-TFBS prediction models, and the AUC was 0.832∼0.993. Overall, RiceTFtarget is a robust web tool for identifying TF-target pairs, which may substantially accelerate the study of the biological roles of TFs and TF regulatory networks in rice.

Km **Prediction Based on Structural Features**: We explored the mutants of enzymes based on Michaelis constants (*Km*) by deep learning based on structural features[4]. The research to optimizing enzyme structures based on deep learning is one of the forefront research hotspots in the field of bioinformatics worldwide. We noticed that although works have carried out to predict *Km* values or other enzyme kinetic parameters like *Kcat* or *Kcat*/*Km*, mutants type are not considered when processing the datasets from Brenda, hence great numbers of useful data were wasted. Therefore, the motivation for this work is to propose a model that can extract more information from enzyme and substrate representations and achieve better performance compared to previous studies. We investigated the UNIREP and ESM2 to extract structural features and GNNs with substrate features, and then the newly introduced deep neural network model demonstrated superior results in *Km* value prediction. This work provides a promisingly approach for enzyme structure optimization.

Other Research

Network and Energy Efficiency

For my PhD research in University of Michigan, Ann Arbor, I worked on devising systematical methodologies and developing tools for characterizing cellular network performance directly from end users. The tools I developed

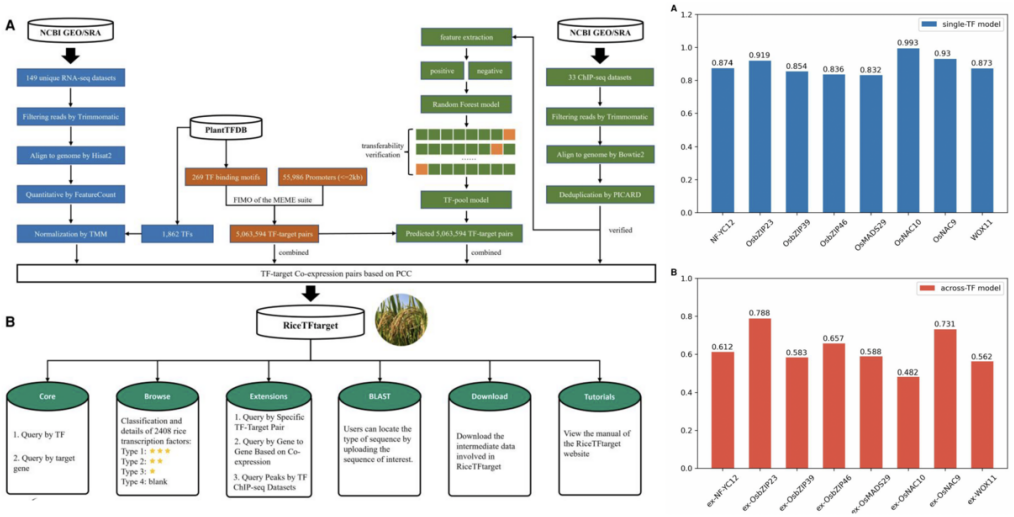


Figure 6: RiceTFTarget – Rice Transcription Factor-Target Prediction Based on Coexpression

as the lead developer includes *3GTest*, *4GTest* and *MobiPerf*, which have cumulatively over 150,000 users from over 190 countries or regions. Notably, *MobiPerf* has received both the *Open Internet App Award* and the *People's Choice App Award* in the *FCC Open Internet Apps Challenge*.

In order to understand the key factors that affect smartphone application performance, I developed a systematic methodology for comparing this performance along several key dimensions such as carrier networks, device capabilities, and server configurations[17]. I performed detailed analysis to help carriers, phone vendors, content providers, and application developers gain insight. I also performed longitudinal study that compare the determinant factors on smartphone applications in 2010[17] and 2012[16]. I identified that the performance bottleneck for web-based applications lies more in the device's processing power than in the network, indicated by the high average CPU usage of 79.3% in LTE network, as well as the underutilized network capacity due to small object size in typical web transactions.

I took one of the first steps in comparing power characteristics of 4G LTE networks with 3G/WiFi networks[16] and developed the first empirically derived comprehensive power model of a commercial LTE network with less than 6% error rate and state transitions matching the specifications. Using a comprehensive data set consisting of 5-month traces of 20 smartphone users, I carefully investigated the energy usage in 3G, LTE, and WiFi networks and evaluated the impact of configuring LTE-related parameters. Despite several new power saving improvements, I found that LTE is as much as 23 times less power efficient compared with WiFi, and even less power efficient than 3G, based on the user traces and the long high power tail was found to be a key contributor.

Despite its fast increasing user base, the network performance and the interplay between mobile applications and the network for the LTE networks still remain unexplored. I thoroughly studied the network performance of LTE network with a data set covering around 300,000 LTE users in a large metropolitan area for 10 days[13]. I revisit basic network metrics in the LTE network and compare with previously studied network conditions. I also observe that a high downstream queueing delay, likely due to bufferbloat, has caused TCP congestion window collapse upon one packet loss. With the help of TCP Timestamps option, I have devised a lightweight passive bandwidth estimation algorithm, allowing us to observe that for 71.26% of the large flows, the bandwidth utilization ratio is below 50%. I find that TCP may not fully utilize the fast-varying available bandwidth when RTT is large in the LTE network. Upon further analysis, I identify 52.61% of all downlink TCP flows have been throttled by TCP receive window and data transfer patterns for some popular applications that are both energy and network unfriendly.

Anomaly Detection with Machine Learning

I've also worked on a number of machine learning projects for anomaly detection.

In Microsoft Research Silicon Valley, we designed and implemented Souche[15], a system that recognizes

legitimate users early in online services by leveraging social connections established over time. Legitimate users help identify other legitimate users through an implicit vouching process, strategically controlled within vouching trees. In our evaluation on a real dataset of several hundred million users, Souche can efficiently identify 85% of legitimate users early, while reducing the percentage of falsely admitted malicious users from 44% to 2.4%. Our evaluation further indicates that Souche is robust in the presence of compromised accounts. It is generally applicable to enhance usability and security for a wide class of online services.

I also designed and implemented a novel framework, called SocialWatch[14], to detect online service abuse attacks at a large scale. We explored a set of social graph properties, ranging from those that describe individual user behaviors, to those that capture the interactions among users and their social affinities. We evaluated SocialWatch using a large dataset from Hotmail with more than 682 million users and over 5.75 billion directional relationships. SocialWatch successfully detects 56.85 million attacker-created accounts with a low false detection rate of 0.75% and a low false negative rate of 0.61%. In addition, this work also addresses the challenge of identifying hijacked accounts within the legitimate account set through a Bayesian decision framework.

Back in 2007, when I was an undergraduate in Tsinghua University, China, I designed and implemented a gesture detection system that uses human gesture patterns to detect deception[18]. The gesture movements in the captured videos are tracked using an efficient computer vision algorithm and the deception behavior is inferred accordingly with the help of machine learning.

Future Work

As shown in my previous research, I have a broad interest in data-driven machine learning approaches originating from real applications. While making new theoretical advances in machine learning is among the top goals, demonstrating the practical impact is also of vital importance. A central thread of my research focuses on devising new machine learning techniques to solve real challenging problems arising from the traditional fields, such as agriculture and manufacturing industry, with software and hardware tools built for helping domain experts and practitioners as the output goals. These tools may help us collect extensive domain data enabling further research work. I plan to extend the scope of my research in the following directions.

Weakly Supervised Learning

During my prior studies, I have observed that in many traditional areas, due to the lack of standardization, public dataset with high-quality labeling is scarce. One example is my work on chicken vocalizations, after extensive search online, we only found very few public datasets available with poor labeling quality and very limited types of audio samples, far from being enough to train a practical model. While it is easy to collect a large amount of data samples with the audio recording system deployed, labeling them accurately is not straightforward. For a significant portion of audio samples, even domain experts may not easily reach consensus on the correct labeling. Another example is our work in defect detection, a nonnegligible number of product images containing defects were falsely labeled to be qualified ones due to various human factors, as many defects are not obvious to observe or to determine, however, they are not tolerable in real production. Moving forward, I am interested in developing new learning algorithms and frameworks with inadequate and incorrect supervision, that can uncover the hidden patterns in the inherent structures of input data, with minimal human annotations required, and allow for iterative improvements by instilling domain knowledge and correcting inconsistent annotations within the existing training data.

Few-Shot Learning

Many real-world applications involve inherent difficulty in generating a large amount of labeled data for training, due to cost and time concerns. For bioinformatics related studies, the cost and time to synthesize a large number of protein sequences and test their functions (enzyme catalytic activity, antibacterial activity, *etc.*) is significant. Sometimes, it may even take several years for researchers in biology and agriculture to carry out phenotype data collection tasks on plants and animals with different genotypes. In manufacturing industry, many new products are produced in small batches according to the market requirement in an extremely short time window, leaving limited time for collecting defective product samples and preparing defect detection model for the new products. Although it might be easy for a human worker to master how to examine a new type of product, it still

remains a nontrivial task for machine learning, without sufficient training data set. Similarly in bioinformatics, in my prior studies, I observe that by preprocessing inferred information of proteins, such as structural features, and instilling more such biological domain knowledge into the model, the performance of the devised machine learning algorithms can be improved on the same training set. Along this angle, I plan to devise a new few-shot learning paradigm, that takes advantage of the comprehensive background domain knowledge with the help of meta learning and prototypical networks, and generate meaningful unseen data samples based on the limited number of available training samples or even zero samples, with the help of generative models, such as generative adversarial networks and variational autoencoders.

High-Throughput Machine Learning

In many of my explored applications, the throughput metric is of vital importance for the designed machine learning algorithms, which might be the ultimate factor that determines the practical feasibility. In order for wide real-world applications, the hardware resources are sometimes reduced to no more than necessary to save the deployment cost. For example, the edge computing devices used in our chicken vocalization project only have 2GB memory with no GPU acceleration available. In many scenarios, the pursuit of higher throughput is limitless. The design goal for our upcoming automated egg sex differentiating machine is 36,000 eggs per hour, and that for our next generation of milk powder peel off lid detection machine is 600 pieces per minute. In addition, for protein and DNA related bioinformatics analysis, the problem space can easily become overwhelmingly enormous, further inspiring extreme performance. I am interested to devise a set of high-through machine learning algorithms to push the speed limit to a much higher level, while minimizing the negative influences in prediction accuracy.

Summary

Broadly speaking, I have rich experiences in entrepreneurship, industry and academia, preparing my research mindset to be real-world scenario oriented. I am interested in making advances in the theoretical principles of machine learning, and devising useful AI tools/systems/machines to solve real challenging problems arising from various interdisciplinary fields, with limited computing resources, inadequate and even inaccurate data labeling, and strict performance requirements, generating positive social and environmental impact.

References

[1] Junxian Huang, Wenwen Xu, Zixin Chen, Jianfeng Gao, Huanliang Xu, Chengming Ji, Zemin Liu, Wei Dai, and Wencai Wu. Method for Detecting Parameters of Chicken Head Organs. *Chinese Patent CN116596937B*, 2023/09/22.

[2] Junxian Huang, Zixin Chen, Jianfeng Gao, Wencai Wu, Wei Dai, Huanliang Xu, Chengming Ji, Meixuan Shan, Wenwen Xu, Zemin Liu, and Jianhua Deng. Method for Detecting Feather Area Parameters of Poultry. *Chinese Patent CN116309791B*, 2023/10/27.

[3] Ming Zhang, Junxian Huang, Jingting Su, Zixin Chen, Xiaojun Ju, Jianfeng Gao, Yunjie Tu, Yanju Shan, Gaige Ji, and Yifan Liu. Method, Device and Equipment for Identifying Characteristics of Pores of Poultry. *Chinese Patent CN116228734B*, 2023/09/22.

[4] Junxian Huang, Tao Yin, Jianfeng Gao, Huanliang Xu, Chengming Ji, Zixin Chen, Wenwen Xu, Zemin Liu. Biological Information Acquisition Method Based on GNN Neural Network. *Chinese Patent CN116312744B*, 2023/09/22.

[5] Junxian Huang, Chengming Ji, Jianfeng Gao, and Zixin Chen, *et al.* A Method for Nondestructive Identification of Chicken Egg Gender Using Hyperspectral Imaging. *Chinese Patent Application*, in Submission, 2024.

[6] Ming Zhang, Junxian Huang, Jingting Shu, Liuchao Zhu, and Xiaojun Ju, *et al.* A Method for Calculating Egg Shape Index and Identifying Breed of Chicken Eggs. *Chinese Patent Application*, in Submission, 2024.

[7] Junxian Huang, Lianzeng Wang, Huaxin Qiao, Shouchang Zhou, and Zixin Chen, *et al.* An Intelligent Health Monitoring System and Method for Poultry Farming. *Chinese Patent Application*, in Submission, 2024.

[8] Lianzeng Wang, Junxian Huang, Shouchang Zhou, and Huaxin Qiao, *et al.* A Method of Diagnosing Respiratory Diseases in Poultry Based on Vocalizations. *Chinese Patent Application*, in Submission, 2024.

[9] Jingting Su, Ming Zhang, Junxian Huang, Yanju Shan, and Zhiwu Chen, *et al.* Method and Device for Poultry Chicken Breed Identification. *Chinese Patent Application CN117593764A*, 2024/02/23.

[10] Baoyi Zhang, Xueai Zhu, Zixin Chen, Hongsheng Zhang, Junxian Huang, and Ji Huang. RiceTFtarget: A Rice Transcription Factor-Target Prediction Server Based on Coexpression and Machine Learning. *Plant Physiology, Volume 193, Issue 1, Pages 190-194*, September 2023.

[11] Zixin Chen, Chengming Ji, Wenwen Xu, Jianfeng Gao, Ji Huang, Huanliang Xu, Guoliang Qian, and Junxian Huang. UniAMP: Enhancing AMP Prediction using Deep Neural Networks with Inferred Information of Peptides. In Submission, 2024.

[12] Chengming Ji, Ke Song, Zixin Chen, Shanyong Wang, Huanliang Xu, Kang Tu, Leiqing Pan, and Junxian Huang. Nondestructive In-Ovo Sexing Identification of Hy-Line Sonia Eggs by EggFormer using Hyperspectral Imaging. In Submission, 2024.

[13] Junxian Huang, Feng Qian, Yihua Guo, Yuanyuan Zhou, Qiang Xu, Z. Morley Mao, Subhabrata Sen and Oliver Spatscheck. An In-depth Study of LTE: Effect of Network Protocol and Application Behavior on Performance. *ACM SIGCOMM*, 2013.

[14] Junxian Huang, Yinglian Xie, Fang Yu, Qifa Ke, Martin Abadi, Eliot Gillum and Z. Morley Mao. SocialWatch: Detection of Online Service Abuse via Large-Scale Social Graphs. *ACM ASIA Conference on Computer and Communications Security (ASIACCS)*, 2013.

[15] Yinglian Xie, Fang Yu, Qifa Ke, Martin Abadi, Eliot Gillum, Krish Vitaldevaria, Jason Walter, Junxian Huang, and Z. Morley Mao. Innocent by Association: Early Recognition of Legitimate Users. *ACM Conference on Computer and Communications Security (CCS)*, 2012.

[16] Junxian Huang, Feng Qian, Alexandre Gerber, Z. Morley Mao, Subhabrata Sen, and Oliver Spatscheck. A Close Examination of Performance and Power Characteristics of 4G LTE Networks. *ACM International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2012.

[17] Junxian Huang, Qiang Xu, Birjodh Tiwana, Z. Morley Mao, Ming Zhang, and Paramvir Bahl. Anatomizing Application Performance Differences on Smartphones. *ACM International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2010.

[18] Fan Xia, Hong Wang, and Junxian Huang. Deception Detection Via Blob Motion Pattern Analysis. *Affective Computing and Intelligent Interaction (ACII)*, 2007.