New Course EFD Template

PURDUE UNIVERSITY | College of Engineering

Engineering Faculty Document
No.: 68-26
October 23, 2025

TO: The Engineering Faculty

FROM: The Faculty of the Agricultural and Biological Engineering Department

RE: New graduate course – **ABE 54200 – Applied ML Systems: On-Device to Cloud Computing**

The Faculty of the Agricultural and Biological Engineering Department has approved the following new graduate course. This action is now submitted to the Engineering Faculty with a recommendation for approval.

FROM (IF ALREADY OFFERED WITH TEMPORARY NUMBER)**:**

ABE 59100

Spring

3.0 total credits; LEC 75/2/16

STAT 30100 OR ABE 20500 OR CHE 32000 OR Graduate standing

Enrollment for last 3 spring offerings was between 10 and 15 students. The first spring it was taught as a 1-credit course and there were 25 students enrolled.

TO:

ABE 54200 - APPLIED ML SYSTEMS: ON-DEVICE TO CLOUD COMPUTING

Spring

3.0 total credits; LEC/75/2/16

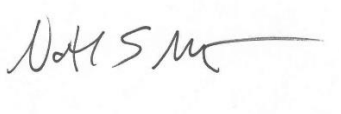
STAT 30100 OR ABE 20500 OR CHE 32000 OR Graduate standing

Design, deploy, and scale on device, vision driven machine learning (ML) systems for the Internet of Small Things (IoST) using the cloud as a selective backstop and foundation model "oracles" for escalation and distillation. The course emphasizes quantization, pruning and distillation, and resource-aware scheduling under tight latency and energy budgets. The course prioritizes on-device computer vision; and uses the cloud when it demonstrably adds value (accuracy, responsiveness, efficiency, or safety). Emphasis is on design reasoning, analysis, and case studies drawn from recent work at the Conference on Neural Information Processing Systems (NeurIPS), the Conference on Computer Vision and Pattern Recognition (CVPR), the International Conference on Computer Vision (ICCV) and the European Conference on Computer Vision (ECCV), the International Conference on Machine Learning (ICML), Operating Systems Design

and Implementation (OSDI), and the European Conference on Computer Systems (EuroSys). Students analyze design choices for end-to-end pipelines targeting mobile Graphics Processing Units (GPUs): supervised and unsupervised learning for detection and segmentation, streaming inference, and minimal cloud backstops for data, evaluation, and safe updates. We treat decentralized learning, with federated learning as a special case, and event-driven streaming in the presence of non-identical data across devices and intermittent connectivity. Evaluation is end-to-end: Intersection over Union (IoU), mean average precision (mAP) across thresholds, 95th percentile latency, memory footprint, and energy per frame, so students learn to navigate accuracy--efficiency tradeoffs. An ethics and impact module examines privacy, safety critical deployment, and sustainability, including the energy footprint of data centers and the role of foundation models as "oracles" and distillers. Use driven, i.e., computing + X, with examples from self-driving cars, mobility, agriculture, and One Health.

RATIONALE:

This will serve as a building block for students wishing to explore further applications of Machine Learning in applications. Theory, foundations, and latest advances in applying machine learning to building computer systems, especially systems under constrained computational resources as in Internet of Things (IoT) devices. In addition to core CV/IoT (computer vision/Internet-of-Things) examples, the course includes use cases in edge computing for digital agriculture (e.g., plant-disease segmentation) and interpretable computing for computational genomics (single-cell genomics clustering), reinforcing the use-inspired, applied ML-systems focus. There have been 5 offerings of the course – the first one was a one-credit offering and the other three were three credit offerings. There were 25 students in the first offering and between 10 and 15 in the others.



Head/Director of the Agricultural and Biological Engineering Department

Link to Curriculog entry: https://purdue.curriculog.com/proposal:34618/form

Syllabus for the Spring 2024 cohort

1/2 of lectures = data science/data engineering/ML concepts and intuition

1/4 of lectures = computer vision and IoT algorithms

1/8 of lectures = reading technical papers and presenting

1/8 of lectures = guest lectures and discussion

Example lectures: https://schaterji.io/teaching/

Brief Course description: This course will cover data science and data engineering algorithms for use in computer vision tasks and IoT devices. This will allow students to develop an algorithmic understanding of the tools and techniques in this field. Further, the course will also include knowledge of commercial software by cloud computing and machine intelligence software vendors such as Google, Microsoft, and Amazon. The course will also cover cloud computing and serverless infrastructure, data engineering best practices, and machine learning innovations targeted at making large networks function on diverse hardware. Finally, the course will also introduce decentralized learning, computational complexity, and ethics to be able to understand the nuances of deploying algorithms in the real world.

Course description for the catalog:

Design, deploy, and scale on device, vision driven machine learning (ML) systems for the Internet of Small Things (IoST) using the cloud as a selective backstop and foundation model "oracles" for escalation and distillation. The course emphasizes quantization, pruning and distillation, and resource-aware scheduling under tight latency and energy budgets. The course prioritizes on-device computer vision; and uses the cloud when it demonstrably adds value (accuracy, responsiveness, efficiency, or safety). Emphasis is on design reasoning, analysis, and case studies drawn from recent work at the Conference on Neural Information Processing Systems (NeurIPS), the Conference on Computer Vision and Pattern Recognition (CVPR), the International Conference on Computer Vision (ICCV) and the European Conference on Computer Vision (ECCV), the International Conference on Machine Learning (ICML), Operating Systems Design and Implementation (OSDI), and the European Conference on Computer Systems (EuroSys). Students analyze design choices for end-to-end pipelines targeting mobile Graphics Processing Units (GPUs): supervised and unsupervised learning for detection and segmentation, streaming inference, and minimal cloud backstops for data, evaluation, and safe updates. We treat decentralized learning, with federated learning as a special case, and event-driven streaming in the presence of non-identical data across devices and intermittent connectivity. Evaluation is end-to-end: Intersection over Union (IoU), mean average precision (mAP) across thresholds, 95th percentile latency, memory footprint, and energy per frame, so students learn to navigate accuracy--efficiency tradeoffs. An ethics and impact module examines privacy, safety critical deployment, and sustainability, including the energy footprint of data centers and the role of foundation models as "oracles" and distillers. Use driven, i.e., computing + X, with examples from self-driving cars, mobility, agriculture, and One Health.

Syllabus

Week-wise Breakdown:

Week 1: Introduction to Machine Learning and Internet of Things
- Internet of Things (IoT) and computer vision (CV) applications -- sensing to actuation
- Machine learning (ML) overview -- supervised, unsupervised, tasks and data lifecycle
- Datasets -- train/validation/test splits, leakage pitfalls
- Example datasets (MNIST, CIFAR, etc.)
- Edge data realities -- label scarcity, imbalance, longtail events

Week 2: Training ML Models and Error Analysis
• Training loops -- loss functions, optimizers, early stopping
• Overfitting vs underfitting -- bias/variance intuition
• Crossvalidation -- kfold, leaveoneout
• Regularization -- ridge, lasso, elastic net; hyperparameter tuning
• Evaluation for vision and systems -- Intersection over Union (IoU), mean average precision (mAP) across IoU thresholds 0.50--0.95, 95thpercentile latency, memory, energy per frame

Week 3: Support Vector Machine (light) and Deep Learning Basics
• Support Vector Machine (SVM) -- when it wins on tiny devices
• Perceptron to modern networks -- capacity vs data
• Deep learning for CV and IoT -- practical pros/cons

Week 4: Activation Functions and Gradient Descent
• Activations -- sigmoid, Rectified Linear Unit (ReLU), softmax (use cases)
• Gradient descent -- batch vs stochastic, learningrate schedules
• Optimization under constraints -- mixed precision, gradient accumulation

Week 5: Information Theory and Model Complexity
• Crossentropy, Kullback–Leibler (KL) divergence, entropy -- why they matter
• Model selection under budgets -- Akaike Information Criterion (AIC) / Bayesian Information Criterion (BIC)

Week 6: Autoencoders and Dimensionality Reduction
• Autoencoders for compression and denoising
• Principal Component Analysis (PCA), tDistributed Stochastic Neighbor Embedding (tSNE), Uniform Manifold Approximation and Projection (UMAP) -- when and why
• Representation compression -- feature distillation for edge

Week 7: Computer Vision Applications (Part 1)
• Detection, segmentation, classification -- core use cases
• Metrics -- accuracy, precision/recall, F1, IoU, mAP across IoU thresholds 0.50--0.95
• Operational constraints -- 95thpercentile latency, frames per second (FPS), memory, energy per frame
• Calibration and thresholds -- confidence tuning for cascades and selective offload

Week 8: Computer Vision Applications (Part 2) -- Transformers and Deployment
• From Convolutional Neural Networks (CNNs) to transformers -- attention without fixed geometry
• Architectures -- Vision Transformer (ViT), Dataefficient Image Transformer (DeiT), Swin Transformer, mobileclass ViTs for ondevice

• Detection/segmentation with transformers -- Detection Transformer (DETR)style heads
• Scale leverage -- maskedautoencoding pretraining, oracle distillation from foundation models, weak labels
• Edgefirst optimizations -- quantization, structured pruning, token reduction, efficient attention, LowRank Adaptation (LoRA)
• Deployment goals -- portable formats; maintain mAP within 1--2 while improving FPS and 95thpercentile latency/memory/energy

Week 9: Machine Learning for Streaming Data
• Streaming for IoT -- realtime ingestion and processing
• Drift and adaptation -- continual learning basics
• Scheduling for streaming -- temporal stride, dynamic resolution and region of interest (ROI), confidencebased cascades
• Selective offload -- bandwidth/privacyaware escalation to cloud or foundationmodel oracles

Week 10: Diffusion Models
• Introduction and relevance to CV/IoT
• Data augmentation and simulation for scarce domains
• Distilling diffusion to lightweight students for edge

Week 11: Cloud Computing Fundamentals
• Cloud for ML/IoT -- patterns and tradeoffs
• Databases -- Structured Query Language (SQL), nonrelational (NoSQL), and object stores for telemetry
• Minimal cloud backstops for the Internet of Small Things (IoST) -- observability, model registry, oracle endpoint

Week 12: Infrastructure for Cloud and IoT
• Serverless workflows -- functions, event triggers, queues
• Edge--cloud orchestration -- backpressure, retries, tail latency
• Deployment patterns -- canary/bluegreen, rollback, cost awareness

Week 13: Responsible ML Systems (Part 1) -- Privacy, Safety, Oracles
• Privacy in IoT/CV -- General Data Protection Regulation (GDPR) basics and data minimization
• Fairness and bias -- measurement and mitigation
• Using foundation models as oracles -- benefits, licensing, privacy risks

Week 14: Responsible ML Systems (Part 2) -- Data Hygiene and Datacenter Impact
• Data provenance and model versioning
• Preprocessing/postprocessing best practices
• Energy and carbon in datacenters -- measuring, mitigating, and ondevicefirst choices

Week 15: Guest Lectures and Review

Examples from past years:

Guest Lecture on Entrepreneurship (Disruptive Innovation)

Guest Lecture on IoT in the Wild (Amazon)

Serverless Cloud Computing (SalesForce)

Diffusion Models (Adobe)

Course Review and Final Discussion

Course Evaluation

- Class participation, pop quizzes, and assessment: 15 points
- Homework: 30 points [method of continuous assessment - e.g., writing summaries, presenting papers and concept, multiple choice/etc]
- Mid-term: 30 points [open-book, open-notes, you can use the same Google doc you create for HWs].

The main purpose of the mid-term will be for you to look up and assimilate concepts, recapitulate the concepts taught and discussed in class, map them to what you learned in class, and then be able to write out the answers formally, uploading your document by the deadline. This is so you get good at technical writing, communicating well to an ML/CS-y audience, capturing the tone of the class. I want this course to whet your appetite to think algorithmically, and to know what to apply, where. This is especially constructive if you are an ML beginner, but also helpful if you are an intermediate (someone who has applied ML packages without a clear intuition) wanting to know the advances in ML at a more algorithmic level. The mid-term exam (and assessment in general) will further this goal.

- End-term: 30 points [open-book, open-notes, open-laptop, needs to be electronically submitted].

I will not accept a hand-written exam to encourage formal writing.

- Instructor feedback: 5 points bonus [by honor code]

This is for the Purdue Qualtrics survey.

Common acronyms

• AIC -- Akaike Information Criterion -- model selection score that penalizes complexity
• AP -- Average Precision -- area under the precision--recall curve at a fixed Intersection over Union threshold
• BIC -- Bayesian Information Criterion -- like AIC but with a stronger penalty on model size
• CNN -- Convolutional Neural Network -- convolutionbased architecture widely used in computer vision
• CV -- Computer Vision -- algorithms that extract understanding from images and video
• DeiT -- Dataefficient Image Transformer -- Vision Transformer variant trained with strong

distillation
• DETR -- Detection Transformer -- endtoend transformer detector using bipartite matching
• F1 -- F1 score -- harmonic mean of precision and recall
• FPS -- Frames Per Second -- throughput metric for realtime systems
• GDPR -- General Data Protection Regulation -- European Union privacy and data protection law
• IoST -- Internet of Small Things -- resourceconstrained, ondevice ecosystems at the edge
• IoT -- Internet of Things -- network of connected sensors/devices that collect and exchange data
• IoU -- Intersection over Union -- overlap metric for localization and segmentation quality
• KL -- Kullback--Leibler divergence -- measure of how one probability distribution differs from another
• LoRA -- LowRank Adaptation -- parameterefficient finetuning using lowrank updates
• MAE -- Masked Autoencoder (in this course) -- selfsupervised pretraining by reconstructing masked patches
• mAP -- mean Average Precision -- AP averaged across classes and multiple IoU thresholds (e.g., 0.50--0.95)
• ML -- Machine Learning -- datadriven modeling and prediction methods
• MLP -- Multilayer Perceptron -- feedforward neural network of fully connected layers
• PCA -- Principal Component Analysis -- linear dimensionality reduction via orthogonal components
• P95 -- 95thpercentile (e.g., latency) -- tail metric capturing worstcase behavior beyond the median
• PTQ -- PostTraining Quantization -- compress a trained model without retraining
• QAT -- QuantizationAware Training -- simulate low precision during training to preserve accuracy
• ReLU -- Rectified Linear Unit -- activation function max(0, x)
• ROI -- Region of Interest -- spatial subset of an image or frame used for focused processing
• SQL -- Structured Query Language -- relational database query and definition language
• NoSQL -- Nonrelational databases -- keyvalue, document, column, or graph stores
• SLO -- Service Level Objective -- target for system performance or reliability (e.g., latency, availability)
• SVM -- Support Vector Machine -- marginbased classifier for linear and kernelized decision boundaries
• Swin -- Swin Transformer -- hierarchical vision transformer with shifted windows for efficiency
• tSNE -- tdistributed Stochastic Neighbor Embedding -- nonlinear visualization of highdimensional data
• UMAP -- Uniform Manifold Approximation and Projection -- fast manifoldpreserving dimensionality reduction
• ViT -- Vision Transformer -- transformer architecture that operates on image patches