Course Title:
Course Number: E301
Credits: 1
Course Directors: Harini Veeraraghavan, PhD veerarah@mskcc.org
Course Prerequisites: None; Open to first year Cancer Engineering PhD students; open to others as approved
Grading Policy: Pass/Fail

Course Description and Learning Objectives

This course provides graduate students in cancer engineering with a comprehensive introduction to statistics, data science, and machine learning as applied to biomedical and clinical problems. Students progress from foundational statistical reasoning through classical machine learning methods to modern deep learning approaches, culminating in hands-on construction of a convolutional neural network for medical image classification. No prior machine learning experience is required; limited programming experience is supported through structured Jupyter notebook assignments with provided starter code.

Upon successful completion of this course, students will be able to:

- **Statistical Analysis:** Analyze data distributions, apply appropriate hypothesis tests, and interpret p-values and confidence intervals in a biomedical context.
- **Performance Metrics:** Calculate and interpret key metrics for evaluating machine learning models, including ROC curves, AUC, precision, recall, and F1-score.
- **Supervised Learning:** Implement and compare supervised classification methods — logistic regression, LASSO, support vector machines, and random forests — on biomedical data.
- **Unsupervised Learning:** Apply clustering (k-means, hierarchical) and dimensionality reduction (PCA, t-SNE, UMAP) to explore high-dimensional datasets.
- **Deep Learning Foundations:** Explain the principles of neural networks, backpropagation, and gradient-based optimization.
- **Convolutional Neural Networks:** Build, train, and evaluate a CNN for classification and understand segmentation architectures in medical imaging.
- **Critical Evaluation:** Assess the limitations, ethical considerations, and failure modes of machine learning models applied to clinical problems.
- **Practical Implementation:** Use Python and Jupyter notebooks to implement end-to-end data analysis and machine learning pipelines.

Course Structure

The course is organized into four thematic parts, building systematically from statistical foundations to modern deep learning. Each week consists of two 90-minute lectures. Jupyter notebook assignments are timed to reinforce the preceding lectures.

- **Part I — Statistical Foundations (Week 1):**Probability and distributions, estimators, hypothesis testing, bootstrapping, and machine learning performance metrics.

- **Part II — Classical Machine Learning (Weeks 2–3):**Supervised learning (logistic regression, LASSO, SVM, random forests) and unsupervised methods (clustering and dimensionality reduction).
- **Part III — Deep Learning (Weeks 4–5):**Neural network fundamentals, CNNs for classification and segmentation, model training best practices, and a hands-on MedMNIST project.
- **Bonus Lecture — Advanced Topics (Week 5, Lecture 10):**High-level overview of transformers, generative models, and foundation models in biomedicine (introductory, no prerequisite knowledge assumed).

Teaching Fellows

Teaching Fellows, drawn from senior GSK students and the postdoctoral community at MSK, are present in the course sessions. Their role is to act as an additional source of information/assistance, to help keep the discussion sessions moving.

Assignments and Methods for Assessing Student Achievement

This course is graded on a pass/fail basis; class participation, assignments and a final project will form the basis of the grading.

Course Evaluation

Students are expected to complete surveys regarding the lectures and overall course via their student portal. This feedback will be used to evaluate the effectiveness and relevance of the topics and provide direction for the subsequent iterations of the course.

Academic Dishonesty, Plagiarism and Artificial Intelligence
The Policy can be found in the Student and Faculty Handbook linked on the GSK Website.

Attendance

Regular attendance is expected. Given the intensive 5-week format, each lecture builds directly on the previous. Students who must miss a class should notify the instructor in advance. Lecture slides will be posted online after each session.

Late Assignments

Assignments submitted up to 48 hours late will receive a 10% deduction. Assignments more than 48 hours late will not be accepted without prior arrangement. Extensions must be requested before the due date.

Programming & Computing

All assignments use Python with Jupyter notebooks. Starter code is provided for all assignments. Students are expected to complete and extend the provided code rather than write from scratch. Google Colab access is available for students without a local GPU. Institutional computing resources (if available) may be used for deep learning assignments.

Use of AI Tools

Students may use AI coding assistants (e.g., GitHub Copilot, ChatGPT) to assist with code completion and debugging. However, all written interpretations, analyses, and conceptual responses must be the student's own work. Submitted work must be understood and explainable by the student.

Academic Integrity

Students are expected to adhere to the institution's Code of Academic Integrity. Collaboration on conceptual understanding is encouraged; however, submitted code and written responses must be independently produced. Plagiarism of any kind will result in a failing grade.

Data Privacy

Only de-identified datasets are used in this course. Students must not share or distribute any course datasets outside the class. Open datasets (MedMNIST, TCIA subsets, or instructor-provided data) will be used for all assignments.

Course Schedule

The course schedule can be found on the following pages and on Moodle.

Weekly Lecture Schedule

Week	Lecture	Title & Topics	Key Concepts	Assignment
1	L1	**Statistical Foundations I: Distributions & Estimation**	• Data types and distributions (normal, binomial, Poisson) • Descriptive statistics: mean, variance, standard deviation • Samples vs. populations; estimators and bias • Frequentist vs. Bayesian perspectives • Confidence intervals	—
	L2	**Statistical Foundations II: Hypothesis Testing & Performance Metrics**	• Null hypothesis, p-values, Type I/II errors • t-tests, Mann-Whitney U, chi-square tests • Multiple testing correction (Bonferroni, FDR) • Bootstrapping and resampling • ML performance metrics: accuracy, ROC/AUC, precision, recall, F1 • Cross-validation strategies	**HW1 Released: Statistical Analysis Notebook**
2	L3	**Supervised Learning I: Logistic Regression & Regularization**	• Supervised vs. unsupervised paradigms • Maximum Likelihood Estimation (MLE) • Logistic regression: model, loss function, interpretation • Regularization: L1 (LASSO), L2 (Ridge), Elastic Net • Feature selection via LASSO; interpreting coefficients • Train/test splits; overfitting vs. underfitting	**HW1 Due** **HW2 Released: Classical ML Notebook**
	L4	**Supervised Learning II: SVM & Random Forests**	• Support Vector Machines: margin maximization, kernels • Hyperparameter tuning (C, gamma) • Decision trees: splitting criteria, depth control • Random forests: bagging, feature importance • Comparing classifiers on biomedical datasets • Model selection and cross-validation	—
3	L5	**Unsupervised Learning I: Clustering Methods**	• Distance metrics in high-dimensional spaces • k-means clustering: algorithm, initialization, choosing k • Hierarchical clustering: linkage methods, dendrograms • Gaussian Mixture Models (GMM) and soft assignments • HDBSCAN: density-based clustering	**HW2 Due** **HW3 Released: Unsupervised Methods Notebook**

			• Evaluating cluster quality (silhouette score, etc.)	
	L6	**Unsupervised Learning II: Dimensionality Reduction**	• Curse of dimensionality • Principal Component Analysis (PCA): variance explained • t-distributed Stochastic Neighbor Embedding (t-SNE) • Uniform Manifold Approximation and Projection (UMAP) • When and how to use each method • Visualizing high-dimensional cancer data	—
4	L7	**Deep Learning Foundations: Neural Networks & Training**	• Biological inspiration and the perceptron • Multi-layer perceptrons (MLP) and activation functions • Forward pass and loss functions • Backpropagation and automatic differentiation • Gradient descent variants (SGD, Adam, RMSProp) • Learning rate schedules; batch size effects • Overfitting: dropout, weight decay, early stopping	**HW3 Due** **HW4 Released: CNN Classification Notebook**
	L8	**Convolutional Neural Networks: Architecture & Classification**	• Convolution, filters, feature maps • Pooling layers and spatial hierarchy • Classic CNN architectures (LeNet, VGG concept, ResNet idea) • Data augmentation strategies • Transfer learning and fine-tuning • Hands-on: building and training a CNN for MedMNIST classification	—
5	L9	**CNNs for Segmentation & Practical Considerations**	• Semantic vs. instance segmentation • Encoder-decoder architectures (U-Net) • Loss functions for segmentation (Dice, cross-entropy) • Model evaluation: Dice coefficient, IoU, Hausdorff distance • Reliability, calibration, and uncertainty • Ethical considerations in clinical AI deployment – Fairness metrics	**HW4 Due**
	L10	**Advanced Topics: Transformers & Generative Models**	• Attention mechanisms and self-attention (high level) • Vision Transformers (ViT) concept • Introduction to generative models: autoencoders, VAE • Generative Adversarial Networks (GANs): intuition	-

			• Foundation models and pretraining in medical imaging • Emerging applications in cancer engineering	

Jupyter Notebook Assignments

All assignments are delivered as Jupyter notebooks (.ipynb) with structured starter code. Students are expected to complete designated code cells, interpret outputs, and write brief responses to conceptual questions. Notebooks are designed to run in Google Colab or a local Python environment. No prior programming experience beyond the provided templates is required.

Assignment	Title	Released / Due	Core Tasks	Deliverable
HW1	**Statistical Analysis & Performance Metrics**	Released: After L2 Due: Start of L3	• Load and visualize a cancer dataset; plot distributions • Compute descriptive statistics and confidence intervals • Perform t-tests and Mann-Whitney U tests; interpret p-values • Apply multiple testing correction • Compute ROC curve, AUC, precision, recall, F1 for a classifier	Completed notebook + brief written interpretation (~1 page)
HW2	**Classical Machine Learning: Classification on Cancer Data**	Released: After L3 Due: Start of L5	• Prepare features; apply train/test split and cross-validation • Train logistic regression and LASSO; interpret coefficients • Train SVM with linear and RBF kernels; compare performance • Train random forest; examine feature importances • Compare all models using ROC/AUC and confusion matrices	Completed notebook + model comparison table
HW3	**Unsupervised Learning: Clustering & Dimensionality Reduction**	Released: After L5 Due: Start of L7	• Apply k-means and hierarchical clustering; choose optimal k • Visualize clusters with silhouette scores and dendrograms • Reduce data with PCA; interpret variance explained • Compare t-SNE and UMAP visualizations of cancer data • Discuss biological interpretation of identified clusters	Completed notebook + written cluster interpretation
HW4	**Deep Learning: CNN Classification with MedMNIST**	Released: After L7 Due: After L9	• Load and preprocess MedMNIST dataset • Build a simple CNN using provided PyTorch/TF starter code	Completed notebook + brief results report (~1 page)

			• Experiment with learning rate and batch size (provided code) • Add a custom layer or data augmentation step • Evaluate model with accuracy, ROC/AUC, and confusion matrix • Discuss model strengths and limitations	

Recommended Resources

There is no required textbook. The following references are recommended for deeper exploration:

Statistics & Machine Learning

- Introduction to Statistical Learning (James et al.) — free PDF available online
- Pattern Recognition and Machine Learning (Bishop) — Bayesian perspective
- MIT OpenCourseWare: Introduction to Probability and Statistics
- scikit-learn documentation: https://scikit-learn.org

Deep Learning

- Deep Learning (Goodfellow, Bengio, Courville) — free online at deeplearningbook.org
- PyTorch tutorials: https://pytorch.org/tutorials
- MedMNIST dataset: https://medmnist.com
- An Introduction to Image Classification (Toennies, Springer 2024)

Cancer Data & Applications

- The Cancer Genome Atlas (TCGA): https://www.cancer.gov/tcga
- The Cancer Imaging Archive (TCIA): https://www.cancerimagingarchive.net
- pyCERR (MATLAB/Python) for radiomics and treatment planning data

Schedule at a Glance

Week	Lectures	Focus
1	**L1 – L2**	Statistical Foundations: Distributions, Hypothesis Testing, Performance Metrics
2	**L3 – L4**	Supervised Learning: Logistic Regression, LASSO, SVM, Random Forests
3	**L5 – L6**	Unsupervised Learning: Clustering, PCA, t-SNE, UMAP
4	**L7 – L8**	Deep Learning: Neural Networks, CNNs, MedMNIST Classification
5	**L9 – L10**	CNN Segmentation, Practical Considerations + Bonus: Transformers & Generative Models