COMP 448/548: Medical Image Analysis

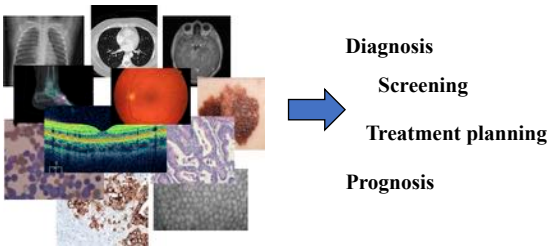
Design pipeline and challenges

Çiğdem Gündüz Demir
cgunduz@ku.edu.tr

1

Last Lecture

- Medical imaging is broad
- There are many medical imaging modalities used in
  - Pathology
  - Radiology
  - Nuclear medicine
  - Ophthalmology
  - Dermatology
  - …



- Each modality works differently to visualize different features of the human body
  - No imaging modality reveals everything
  - Computational analysis tools should be designed taking into account the imaging modality and the manner it is used

2

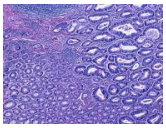
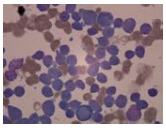
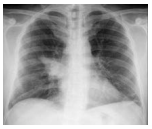
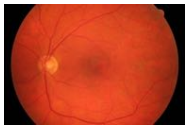
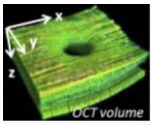
Outline for Today

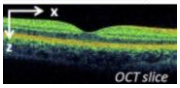
- Preliminaries
- Design pipeline (overall picture)
- Challenges
- Design pipeline (algorithm design)

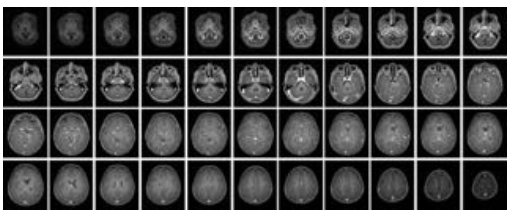
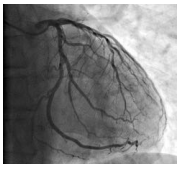
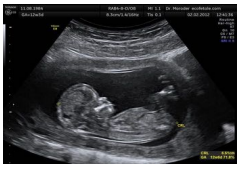
3

What is an input?

- 2D image
  - Histopathology image
  - Blood smear image
  - X-ray image
  - Fundus photograph
  - …
- 3D volume
  - CT scan
  - MR scan
  - OCT
  - …
- Video
  - Angiography
  - Endoscopy
  - …



4

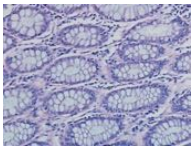
What is a 2D image?

- It is a two-dimensional function f(x, y) that gives the intensity at position (x, y)
- Images are usually discrete
  - Grayscale: f(x, y) → [0, 255]
  - RGB color: f(x, y) is a 3D vector

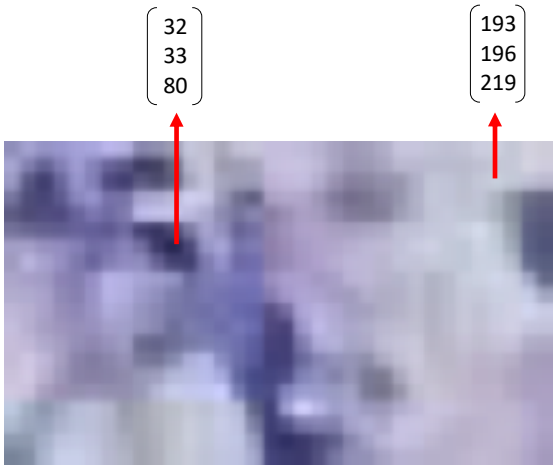
$$f(x,y) = \begin{bmatrix} r(x,y) \\ g(x,y) \\ b(x,y) \end{bmatrix}$$

r(x, y) → [0, 255]
g(x, y) → [0, 255]
b(x, y) → [0, 255]

  - Other color spaces: Lab, HSV, …



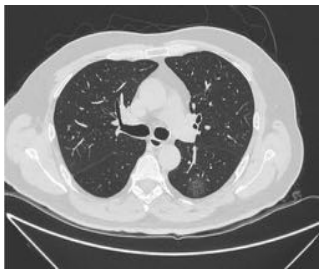
Bright regions: higher intensities
Dark regions: lower intensities



5

What is a 2D image?

- It is a two-dimensional function f(x, y) that gives the intensity at position (x, y)
- Images are usually discrete
  - Grayscale: f(x, y) → [0, 255]
  - RGB color: f(x, y) is a 3D vector
  - Other color spaces: Lab, HSV, …



Bright regions: higher intensities
Dark regions: lower intensities

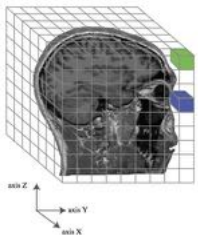
Predefined threshold value	Minimum	Maximum
Bone (CT)	226	3071
Soft Tissue (CT)	-700	225
Enamel (CT, Adult)	1553	2850
Enamel (CT, Child)	2042	3071
Compact Bone (CT, Adult)	662	1988
Compact Bone (CT, Child)	586	2198
Spongial Bone (CT, Adult)	148	661
Spongial Bone (CT, Child)	156	585
Muscle Tissue (CT, Adult)	-5	135
Muscle Tissue (CT, Child)	-25	139
Fat Tissue (CT, Adult)	-205	-51
Fat Tissue (CT, Child)	-212	-72
Skin Tissue (CT, Adult)	-718	-177
Skin Tissue (CT, Child)	-766	-202

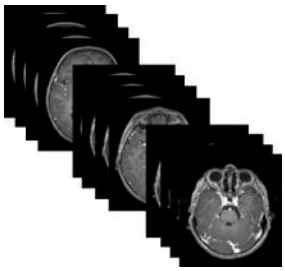
Hounsfield unit (HU) for various organs of human body

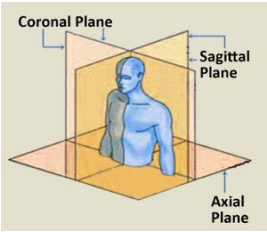
6

What is an input?

- 2D image is a matrix of pixels, each of which has a value of f(x, y)
- 3D volume is a tensor of voxels, each of which has a value of f(x, y, z)





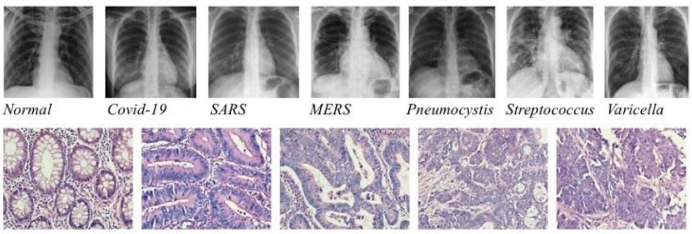


- Video is a sequence of individual video frames (images)

7

What is an output?

- Estimate a single output for an entire image
  - **Classification** when the output is discrete (binary or multiclass)
    - Binary: output is 0 or 1
    - Multiclass: one-hot coded outputs are used for neural networks
      - E.g., if the class labels are 0 to 5, the output for the 3rd class is 0 0 0 1 0 0



Chest X-ray classification

Tissue image classification

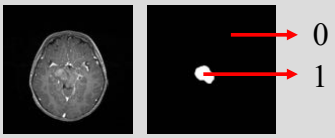
  - **Regression** when the output is continuous
    - E.g., estimating the risk of coronary artery disease, in terms of a continuous value [0, 100]

8

What is an output?

- Estimate an output for every pixel in an image (a map for the entire image)
  - Semantic segmentation when the output of each pixel is discrete

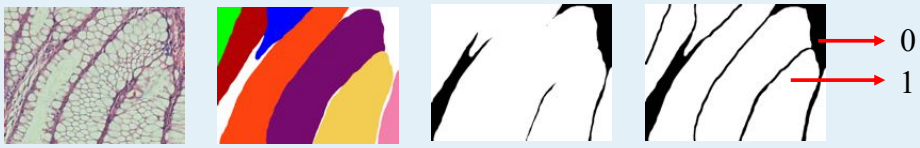
Foreground / background segmentation is very common → *binary classification*



More than two segmentation labels → *multiclass classification*

In instance segmentation, the annotation map may contain different labels for each instance

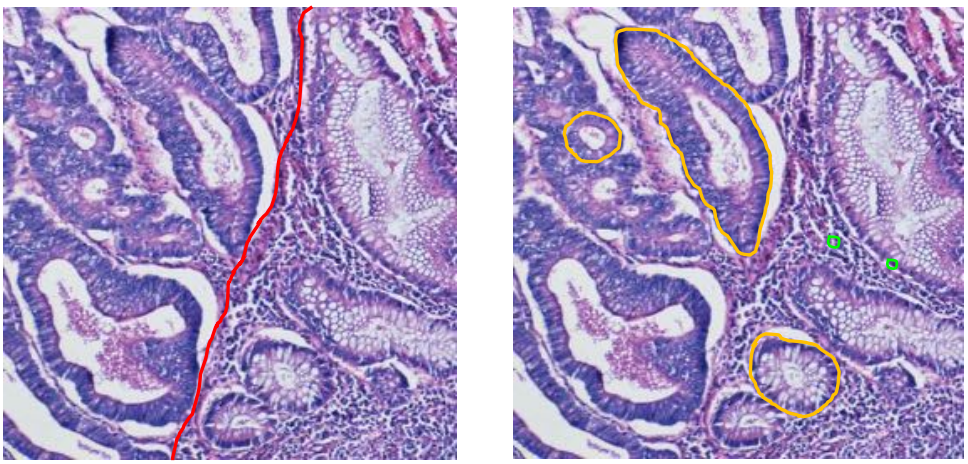
- You need to process these labels → still *binary classification* if all instances belong to the same class



9

What is an output?

- Estimate an output for every pixel in an image (a map for the entire image)
  - Output depends on what you want to focus on in your application



10

What is an output?

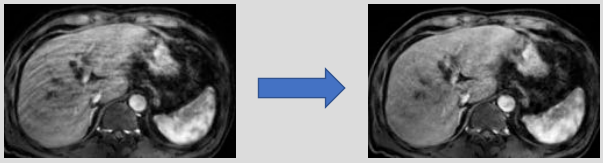
- Estimate an output for every pixel in an image (a map for the entire image)
  - Regression when the output of each pixel is continuous
  - Examples include image reconstruction, image synthesis, and artifact reduction

MR image synthesis from a CT image

Input: CT image whose pixels are in [0, 255]

Output: MR image whose estimated pixels are also in [0, 255]

Artifact reduction in MR images



Input: Raw MR image whose pixels are in [0, 255]

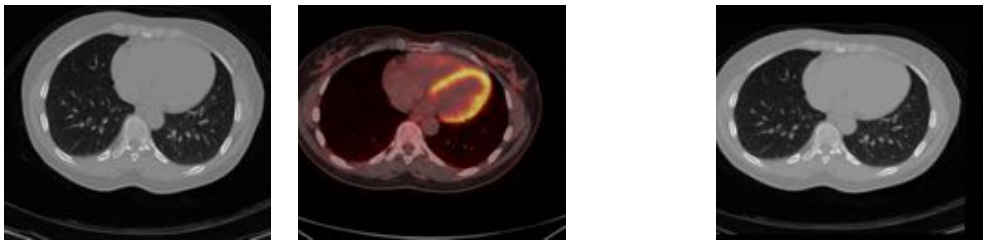
Output: Artifact-reduced MR image whose estimated pixels are also in [0, 255]

- Content-based image retrieval (CBIR) is the problem of finding similar images to a query image in an image dataset → outputs are retrieved images

11

What is an output?

- **Image registration** is the process of transforming different sets of data into one coordinate system



12

Design pipeline

Problem definition and dataset preparation

- Define a problem
- Collect data
- Annotate data

Build an image analysis model

- Design an algorithm
- Select parameters (if any)
- Train the model (if required)

Evaluate the model

- Evaluate the model performance *visually and quantitatively*
- Analysis *comparison with the existing approaches, ablation studies, and parameter analysis*

13

Challenges

Prob. definition and dataset preparation
- **Define a problem**
- Collect data
- Annotate data

- Defining a "*meaningful*" problem is usually hard, regardless of the domain
- But, it is typically harder in the domain of medical image analysis
  - As a computer scientist or an engineer, you may not be that familiar to the problem domain
  - As a clinician or a biologist, you may not know the capabilities and limitations of computer algorithms
  - All parties need to work in close collaboration and need to speak a common language

14

Challenges

Prob. definition and dataset preparation
- Define a problem
- **Collect data**
- Annotate data

- Data collection takes effort and time
  - Ethics review committee approvals are necessary for data collection (both for studies involving human and lab animal subjects)
  - Informed consent should be sought from human subjects
- Non-standardized preparation of samples and/or non-standardized acquisition of images
  - When there are more than one data source
  - When images are acquired at different times, with different systems, and from different labs
  - **The more variety there is in the data, the larger the dataset needs to be**

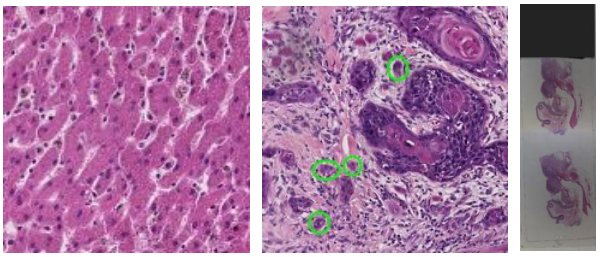
15

Challenges

Prob. definition and dataset preparation
- Define a problem
- Collect data
- **Annotate data**

- Annotation is very challenging
- It needs medical expertise
- There might be inconsistencies in annotations
  - Sometimes there is no consensus among annotators
    Remember intra and inter-observer variability
  - There may exist hard-to-annotate image parts and incorrect annotations as a result
    - Due to noise and artifacts as well as due to the nature of problem
    - E.g., Detecting and marking all true positives in an image may not be possible or require too much effort

It is not like preparing a dataset for example for the application of pedestrian detection in a street, for which pedestrians can be marked by almost any person



16

Challenges

Build an image analysis model
- **Design an algorithm**
- Select parameters (if any)
- Train the model (if required)

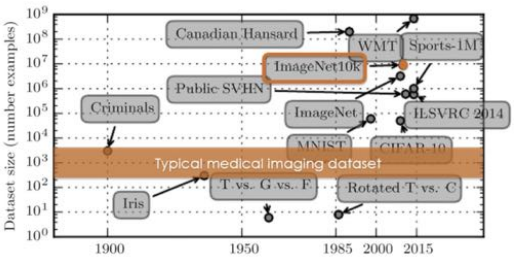
- Hard to define expressive features
  - Large variance may exist within samples
  - Noise and artifacts typically exist in samples/images due to non-ideal conditions in experimental setup and imaging
- Traditionally, features are manually defined based on domain-specific knowledge, human intuition, and known mathematical theories and tools
  - Sometimes, this process of extracting handcrafted features is not "*that effective*"
- Recently, deep learning has shown great promise as an alternative to employing handcrafted features
  - But it requires large datasets for training

17

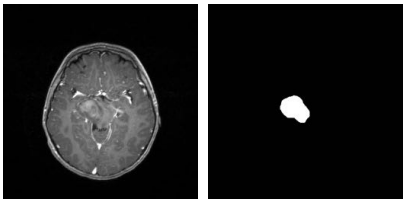
Challenges

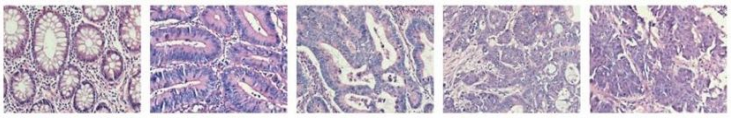
Build an image analysis model
- Design an algorithm
- Select parameters (if any)
- **Train the model (if required)**

- Need of large annotated datasets to train deep models
  - **The more variety there is in the data, the larger the training dataset needs to be**
- Difficult to access to large high-quality annotated datasets
  - ImageNet is extremely powerful since it is huge and accurately annotated
- Imbalanced data problem



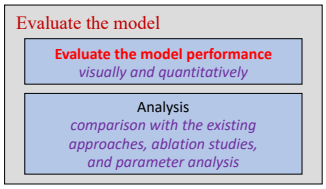
www.quantib.com/blog/deep-learning-radiology-and-challenges-radiology-ai





18

Challenges



- Model is evaluated by estimating its performance
  - Misleading conclusions if it is not done properly
- Bias in the estimate: Performance on the training set samples is often a poor estimator of the performance over future samples
  - Likelihood of a model to overfit the training set samples is high especially when the model is complex, parameters are finetuned, and the training set is small
  - To obtain an unbiased estimate of the future performance, the model should be tested on samples chosen independently of the training samples and the model
- Variance in the estimate: The measured performance can vary from the true performance depending on the test set samples
  - The expected variance is high especially when the test set is small

19

Estimating the classification error

- Given a model M and a dataset S containing n samples drawn at random according to a distribution D
  1. What is the best estimate of the error of M over future samples drawn from the same distribution?
  2. What is the probable error in this error estimate?

The **true error** of the model M with respect to a target function f and the distribution D is the probability that M will misclassify a sample drawn at random according to D

$$error_D(M) \equiv \Pr_{x \in D}\left[f(x) \neq M(x)\right]$$

The **sample error** of the model M with respect to f and the dataset S

$$error_S(M) \equiv \frac{1}{n}\sum_{x \in S}\delta(f(x), M(x)) \qquad \delta(f(x), M(x)) = \begin{cases} 1 & \text{if } f(x) \neq M(x) \\ 0 & \text{otherwise} \end{cases}$$

How good an estimate of $error_D(M)$ is provided by $error_S(M)$?

20

Estimating the classification error

- Given no other information, the most probable value of $error_D(M)$ is $error_S(M) = r / n$, where r is the number of misclassified samples, and
- With N percent confidence, $error_D(M)$ lies in the interval of

$$error_S(M) \mp z_N \sqrt{\frac{error_S(M)\left(1 - error_S(M)\right)}{n}}$$

z_N should be chosen depending on the desired confidence level

  - If dataset S contains n samples drawn independent of one another, and independent of model M, according to the distribution D, and
  - If $n \geq 30$ [more accurately, $n\, error_S(M)\,(1 - error_S(M)) \geq 5$]

Confidence level N%	50 %	68 %	80 %	90 %	95 %	98 %	99 %
Constant z_N	0.67	1.00	1.28	1.64	1.96	2.33	2.58

21

How to find this confidence interval?

- Binomial distribution
  - Let's consider a coin-tossing experiment to find the probability p of obtaining head
  - This experiment involves n trials, in each of which we obtain either head (1) or tail (0)
    - Each trial is Bernoulli
    - Thus, the entire experiment follows the Binomial distribution

$$P(X = r) = \frac{n!}{r!\,(n-r)!}\, p^r\, (1-p)^{n-r}$$
$$E[X] = n\,p$$
$$Var(X) = n\,p\,(1-p)$$
$$\sigma_X = \sqrt{n\,p\,(1-p)}$$

For sufficiently large values of n, the Binomial distribution is closely approximated by a normal distribution with the same mean and variance

- *Design an experiment to find the probability p of misclassification*
- *This experiment involves classifying the samples of a randomly drawn set with a size of n*
- *For each sample, we obtain either misclassification (1) or correct classification (0)*

22

How to find this confidence interval?

- Derive a confidence interval for $error_D(M)$
  - The derivation is quite tedious for the Binominal distribution
  - If p follows a normal distribution, the measured p will fall the following interval N% of the time $\mu_p \mp z_N \sigma_p$

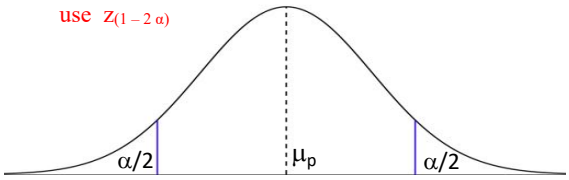
N% confidence interval for p is an interval that is expected to contain p with N% probability

- Thus, $error_D(M)$ will fall the following interval with N% confidence

$$error_S(M) \mp z_N \underbrace{\sqrt{\frac{error_S(M)\left(1-error_S(M)\right)}{n}}}_{\text{Multiplying a random variable by constant } n \text{ multiplies the variance by } n^2}$$

This gives two-sided bounds with N% confidence (α significance level where α = 1 – N%)

For one-sided bound with the same confidence, use $z_{(1-2\alpha)}$



23

What does it mean?

10 misclassifications of 100 samples → $error_S(M) = 0.1$
with 95% confidence, the true error lies in the interval of

$$0.1 \mp \underbrace{1.96\sqrt{\frac{0.1\ 0.9}{100}}}_{\sim 0.059} \Rightarrow 0.041 \le error_D(M) \le 0.159$$

100 misclassifications of 1000 samples → $error_S(M) = 0.1$
with 95% confidence, the true error lies in the interval of

$$0.1 \mp \underbrace{1.96\sqrt{\frac{0.1\ 0.9}{1000}}}_{\sim 0.019} \Rightarrow 0.081 \le error_D(M) \le 0.119$$

10 000 misclassifications of 100 000 samples → $error_S(M) = 0.1$
with 95% confidence, the true error lies in the interval of

$$0.1 \mp \underbrace{1.96\sqrt{\frac{0.1\ 0.9}{100\ 000}}}_{\sim 0.002} \Rightarrow 0.098 \le error_D(M) \le 0.102$$

SO WHAT DOES IT MEAN?

z_N =1.96 for two-sided 95% confidence interval

24

How to evaluate the model performance?

- To obtain an unbiased estimate of the future performance, the model should be tested on samples chosen independently of the training set samples and the model → **TEST SET**

- How to form a test set(s)?
  - One separate test set
  - Multiple test sets

25

How to evaluate the model performance?

- One separate test set
  - If one is available (e.g., if you use a public dataset), use it as it is
  - If not, randomly split the data into two
    - Consider class distributions
    - Consider dependency between the samples (if any)
      - No dependency should exist in the ideal case
      - However, dependency may exist in practice
        (e.g., multiple images from the same biopsy specimen → patient dependent)

26

How to evaluate the model performance?

- Multiple test sets → partition the data many times
  - ***Bootstrapping:*** Draw samples from the dataset with replacement
  - ***K-fold cross validation:*** Form k partitions of the dataset
  - ***Leave-one-out:*** Form partitions, each containing a single sample

- For all, you need to consider the dependency among samples if any
  - E.g., if there are multiple images from the same patient
    - Form your partitions accordingly in k-fold cross validation
    - Do not use leave-one-image-out but use leave-one-patient-out

27

How to select the model parameters?

- Using the correct parameter values is essential for any algorithm
- However, you should not finetune them
- And more importantly, **you should not select them on the test set**
- Then, how to select?

28

Grid search

- Define a search space as a grid of parameter values and evaluate every position in the grid
  - Determine a set of values for each parameter
  - Consider every combination of the selected values in these sets
  - Select the combination that gives the best performance
- Consider this selection as a part of training, but do not use the training set performance as the selection criterion to prevent overfitting
  - You may use a separate validation set or
  - You may use k-fold cross-validation on the training set and select the parameter combination that leads to the highest average cross-validation performance
  - ***Do not use the test set performance in any step of this selection***

29

How to select the model parameters?

- Then, how to select?
  - Grid search defines a search space as a grid of parameter values and evaluates every position in the grid
  - Random search defines a search space as a bounded domain of parameter values and randomly samples points in that search space
  - Bayesian optimization builds a probability model of the objective function and then iteratively uses this model to select the most promising parameter values and updates the model
  - Other optimization methods

30

Analysis

- Ablation studies to understand the contribution of each component of a model to its overall performance
  - Remove each component and keep the other components exactly the same and measure the performance
- Comparisons with the well-known and recent studies in the literature
- Parameter analysis to understand the effects of each parameter to the model performance
  - For each parameter, fix the selected values of the other parameters and measure the model performance as a function of the parameter of interest

31

Is the difference between the performance of two classifiers statistically significant?

- **Use the Mc Nemar's test if you have their results on a single test set**

Accept the hypothesis that two classifiers have the same error rate at a significance level α if

$$\frac{\left(\left|e_{01}-e_{10}\right|-1\right)^2}{e_{01}+e_{10}} \leq \chi^2_{\alpha,1}$$

e_{01}: number of samples misclassified by the first classifier but not by the second one

e_{10}: number of samples misclassified by the second classifier but not by the first one

Example: Are the following two algorithms same with a significance level of 0.05?

	A	B
100	+	+
12	+	−
25	−	+
30	−	−

$$\frac{\left(\left|25-12\right|-1\right)^2}{25+12} = 3.8919$$

Significance level α	0.20	0.10	0.05	0.02
$\chi^2_{\alpha,1}$	1.64	2.71	3.84	5.41

Chi-square statistics with dof = 1

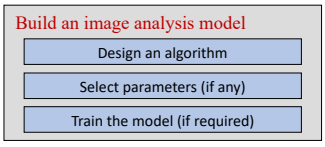
32

Is the difference between the performance of two classifiers statistically significant?

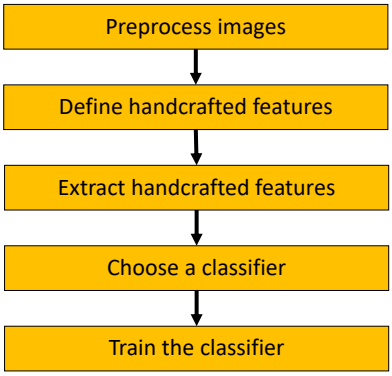
- **If you have their results on multiple test sets**
  - Paired t-test (parametric test)
    - Assumes that the test set errors for both of the classifiers are normally distributed so their differences are
    - Uses the t-test to check whether or not the mean of these differences is equal to zero (statistically significantly)
  - Wilcoxon signed-rank test (nonparametric test)
    - Ranks the differences in errors (ignoring the signs) and sums the ranks for the positive and negative differences (corresponding to the 1st and 2nd classifiers)
    - Claims that the difference between the classifiers is statistically significant if the smaller of the sums is smaller than the critical value defined for the Wilcoxon test

33

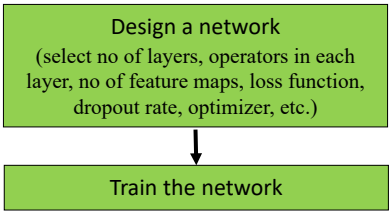
Examples of design pipelines



Traditional approach



Deep learning approach



34

Examples of design pipelines

Build an image analysis model

- Design an algorithm
- Select parameters (if any)
- Train the model (if required)

Traditional approach

Preprocess images
↓
Define handcrafted features
↓
Extract handcrafted features
↓
Design a segmentation algorithm
(e.g., region growing on the similarity of features between adjacent pixels/regions)
↓
Postprocess the segmentation maps

Deep learning approach

Design a network
(select no of layers, operators in each layer, no of feature maps, loss function, dropout rate, optimizer, etc.)
↓
Train the network
↓
Postprocess the segmentation maps

35

Thank you!

Next time:

Filters for image enhancement

36