# Literature Survey: Trends in Medical Image Analysis by Deep Learning


- VIDESH SUMAN (150040095)


R & D PROJECT (EE 691)

GUIDE: PROF. AMIT SETHI

# Introduction

- Computer aided analysis for better interpretation of images
- Early detection, diagnosis, treatment of diseases
- Early 2000s: Shift from systems completely designed by humans, to feature engineered supervised learning systems
- Rise of Deep Learning
- Input to output by learning higher level features in the training process itself
- Better accuracy in all aspects (classification, segmentation, detection, etc.)

# Deep Learning Tasks in Medical Image Analysis

- Classification:
  - Images
  - Objects/Lesions
- Detection:
  - Organ/region Localization
  - Lesion Detection
- Organ/Lesion Segmentation

# A deep-learning classifier identifies patients with clinical heart failure using whole-slide images of H&E tissue (2018)

- JEFFREY J. NIRSCHL, ANDREW JANOWCZYK, ELIOT G. PEYSTER, RENEE FRANK, KENNETH B. MARGULIES, MICHAEL D. FELDMAN, ANANT MADABHUSHI

# Introduction

- Heart Failure – Serious, progressive clinical syndrome
- Diagnosis relies on **clinical history**, physical exam, imaging and EMB (Endomyocardial biopsy)
- However, manual EMB interpretation has **high inter-rater variability**
- Alternative: computer-aided analysis – Deep Learning
- CNN classifier to detect clinical heart failure from sub-images sampled from Whole Slide Images of cardiac tissue
- Outperforms conventional feature-engineering approaches and two expert pathologists

# Dataset Description

- Dataset: left ventricular tissue from **209** patients, collected at the University of Pennsylvania between 2008 and 2013
- The tissue was procured from two separate groups of subjects:
  - heart transplant or LVAD recipients with severe heart failure (Failing or Fal; **94**), and
  - brain dead, organ donors with no history of heart failure (Non- failing or NF; **115**)
- Random split: **104** patients for training, and the rest **105** held out for testing
- At the patient level, training set was split into **3 folds** for cross validation

# Image Analysis Pipeline

- For each patient's WSI, down sampled to **5x magnification**, **11** non-overlapping images or regions of interest (ROI 2500µm2) were extracted at random from within the **Otsu-thresholded** (binarization algorithm) and manually refined tissue border.
- 2 independent classifiers modelled:
  - **Fully-convolutional** CNN architecture very similar to AlexNet, a bit modified to transform the image pixels (input) into the probability (output) that the image came from a patient with heart failure
  - traditional feature-engineering approach using WND-CHARM, a generalized image pattern recognition system coupled to a random decision forest classifier (RF)
- Images with probabilities >50% - "Failing" class at the image-level.
- The image-level predictions were grouped by patient and the fraction of images predicted as 'Failing' for each patient gave the patient- level probabilities

# CNN Model

- Fully-convolutional architecture composed of alternating convolutional, batch normalization, and RelU activation layers
- Approximately 13,500 learnable parameters
- The classifier was trained using **100 patches per ROI, per patient**, and the training set was augmented rotating each patch by 90 degrees
- CNN Output: An image-level probability of whether ROIs belong to the failing class
- Each fold of the three-fold cross validation was trained using NVIDIA DIGITS for **30 epochs** on a Titan X GPU with CUDA7.5 and cuDNN optimized by Stochastic Gradient Descent built into Caffe and a fixed **batch size of 64**

| Layer | Kernel size | Kernels | Stride | Output size |
|---|---|---|---|---|
| Input | | | | 3 x 64 x 64 |
| Conv 1a | 3x3 | 16 | 1 | 16 x 62 x 62 |
| Conv 1b | 2x2 | 16 | 2 | 16 x 31 x 31 |
| Conv 2a | 3x3 | 16 | 1 | 16 x 29 x 29 |
| Conv 2b | 3x3 | 16 | 2 | 16 x 14 x 14 |
| Conv 3a | 3x3 | 16 | 1 | 16 x 12 x 12 |
| Conv 3b | 4x4 | 16 | 2 | 16 x 5 x 5 |
| Fc-conv | 5x5 | 2 | - | 2 |

*Source: Jeffrey J. Nirschl et al.*

- The network accepts **64x 64 pixel RGB** image patches (128x128µm) with a corresponding label.

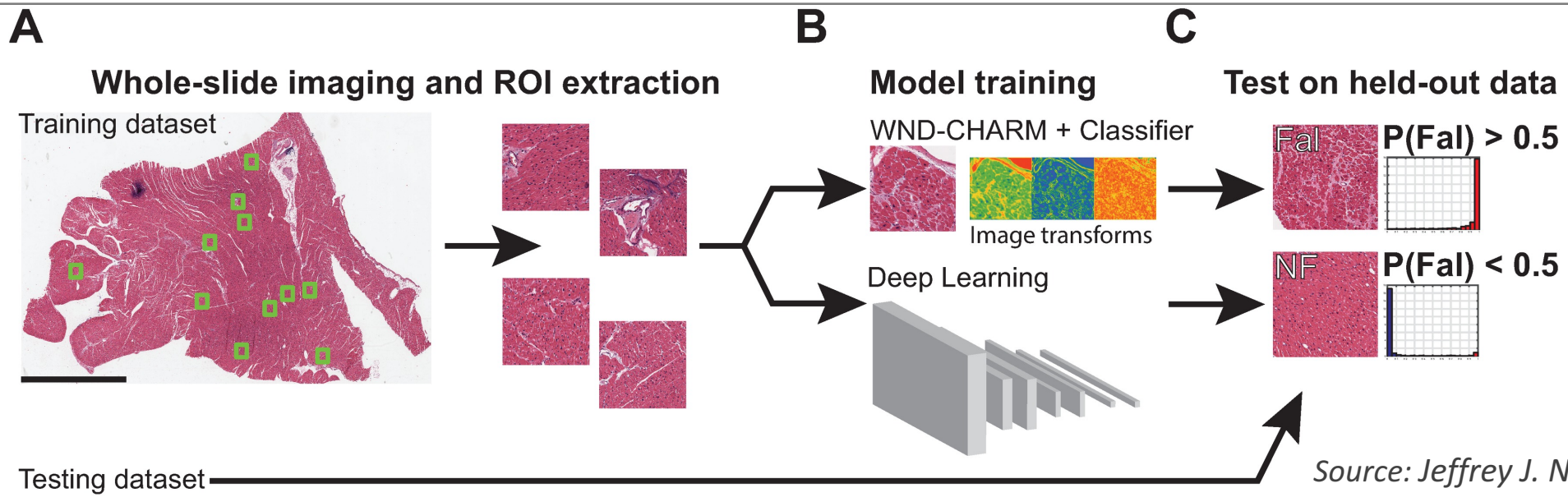
# WND-CHARM + RF Model

- The comparative approach used WND-CHARM to extract **4059 engineered features from each ROI**. This rich feature set was shown to perform as well or better as other feature extraction algorithms on a diverse range of biomedical images
- **The top 20 features** shown alongside were selected using the minimal Redundancy Maximal Relevance algorithm
- mRMR: It tends to select a subset of features having the **most correlation with a class** (relevance) and the **least correlation among themselves** (redundancy).
- These features were used to train a **1000 tree** Breiman-style random decision forest using the TreeBagger function in MATLAB.
- The output of the random decision forest was an image-level **probability of whether an ROI belongs to the failing class**.

1. Color Histogram [17]
2. Multiscale Histograms [20]
3. Haralick Textures (Hue()) [9]
4. Pixel Intensity Statistics (Chebyshev (Hue())) [2]
5. Multiscale Histograms (Chebyshev (Fourier ())) [9]
6. Zernike Coefficients (Wavelet ()) [19]
7. Color Histogram [4]
8. Zernike Coefficients (Fourier ()) [22]
9. Pixel Intensity Statistics [4]
10. Chebyshev Coefficients (Fourier (Edge ())) [8]
11. Color Histogram [5]
12. Chebyshev-Fourier Coefficients (Fourier (Hue())) [12]
13. Gabor Textures [4]
14. Pixel Intensity Statistics (Edge ()) [4]
15. Zernike Coefficients (Fourier ( Edge ())) [9]
16. Zernike Coefficients (Fourier ( Edge ())) [17]
17. Pixel Intensity Statistics [3]
18. Comb Moments (Hues ()) [44]
19. Color Histograms [3]
20. Comb Moments (Wavelet (Fourier ())) [9]

*Source: Jeffrey J. Nirschl et al.*

# Schematic overview of digital pathology workflow to detect heart failure




*Source: Jeffrey J. Nirschl et al.*

(a) Patients were divided into a training and test dataset. WSI were scanned and regions of interest (ROI) were extracted for image analysis. All ROIs from the same patient were given the same label, which was determined by whether the patient had clinical or pathological evidence of heart disease. (b) Three-fold cross validation was used to train heart failure classifiers using a deep learning model & engineered features in WND-CHARM + a random forest classifier. (c) Trained models were evaluated at the image and patient-level on a held-out test dataset.

# Results: CNN identifies heart failure patients with high accuracy

- CNN performed very well, detecting heart failure at both the image and patient-level with accuracy exceeding **93%** on both the training and test datasets
- Though the WND-CHARM + RF model was able to identify heart failure patients, the accuracy and sensitivity was **significantly lower** than the CNN model at both the image and patient-level on the held-out test set
- The learned features in the CNN result in a classifier that is **more discriminative** than a feature-engineering approach, which is capable of accurately detecting heart failure from cardiac histopathology
- Note:
  - Sensitivity: True Positives / Actual Positives
  - Specificity: True Negatives / Actual Negatives
  - where, Positive: Failing class

# Results: Comparison with the human performance

- In the clinical setting, pathologists don't take such decisions by only assessing the tissue images
- However, in order to assess the human performance at this task, two expert pathologists were trained on **this training set of 104 patients** (11 images per patient along with the ground truth)
- Then, they independently reviewed **105 patients in the test set** with no time constraints
- For each set of patient images, they gave a **binary prediction** of whether the set of 11 images were from a patient with clinical heart failure or not
- Both had an **individual accuracy of 75%** at the patient-level with a **Cohen's kappa inter-rater agreement of 0.40**.
- The CNN significantly outperformed pathologists in all metrics with a **20%** differential in terms of sensitivity and specificity
- Though this task required pathologists to perform a **non-standard task** that **differs from their standard diagnostic workflow**, this format ensured that the **algorithms and pathologists were given the same set of information** to make their respective predictions

# Results: Image and patient-level performance evaluation

| Held-out test | Random Forest | Deep Learning | Pathologists |
|---|---|---|---|
| | | | |
| **Image-level** | | | |
| Accuracy | 0.862 ± 0.01 | 0.932 ± 0.01 | - |
| Sensitivity | 0.909 ± 0.02 | 0.985 ± 0.01 | - |
| Specificity | 0.823 ± 0.03 | 0.900 ± 0.002 | - |
| Positive Predictive Value | 0.810 ± 0.03 | 0.878 ± 0.003 | - |
| **Patient-level** | | | |
| Accuracy | 0.895 ± 0.03 | 0.940 ± 0.03 | 0.75 0.75 |
| Sensitivity | 0.979 ± 0.01 | 1.00 ± 0.001 | 0.81 0.64 |
| Specificity | 0.828 ± 0.05 | 0.891 ± 0.01 | 0.71 0.85 |
| Positive Predictive Value | 0.823 ± 0.04 | 0.881 ± 0.01 | 0.69 0.77 |

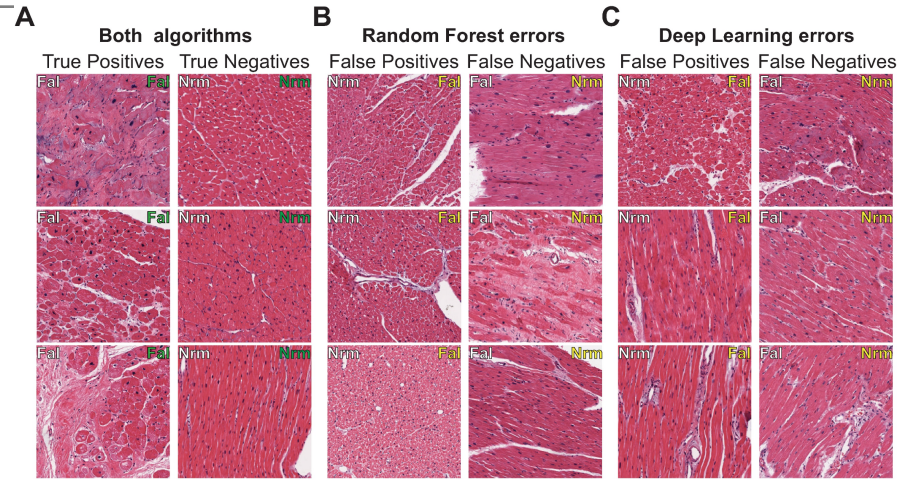
- Note:

*Source: Jeffrey J. Nirschl et al.*

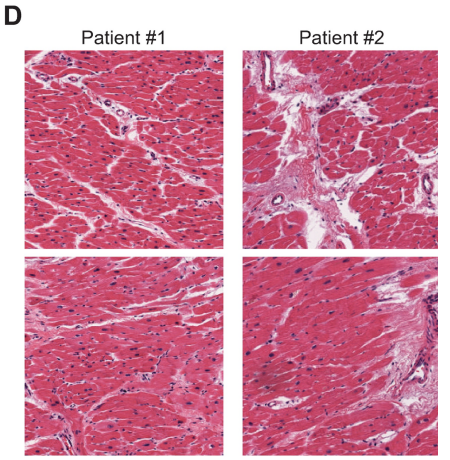
Positive Predictive Value: True Positive / Total Predicted Positive

# Results: Review of correctly and incorrectly classified images

- Qualitative review of these images by expert pathologists showed that a few images have **some features of both normal and abnormal myocardium**.
- This could make these images **challenging for both computers and humans to classify** when only given a single image.



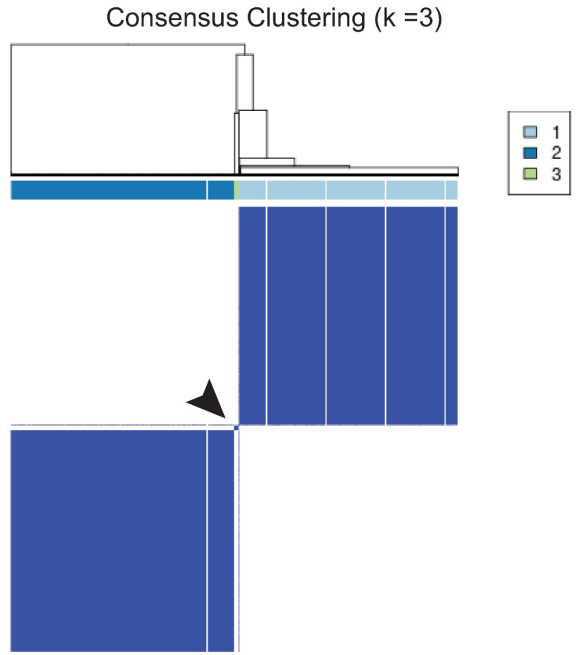

- It was found that several non-failing patients were unanimously **misclassified with a very high probability** at both the image and patient-level. This prompted a **detailed re-examination** of these cases. Two expert pathologists on reviewing such cases independently, reached a consensus agreement that **two patients** exhibited **severe tissue-level pathology** (Fig. D) while the other cases had mild to moderate pathology or tissue-processing artifacts, with **no pre-existing heart failure**.




*Source: Jeffrey J. Nirschl et al.*

# Results: Review of abnormal cases and unsupervised clustering of patients

- Although many patients clustered with patients that had similar heart function, a **small subset of patients** fell into a third, smaller cluster in-between the failing and normal clusters.
- Misclassified patients without heart failure, but who had severe tissue pathology, either fell into this third cluster or had a majority of their images in the failing cluster.
- This prompted the designation of a **new ground truth label** based on both the clinical history and histopathological findings. The two non-failing patients described above were **reassigned** to an "abnormal or heart failure" class due to their severe pathology and reproducible clustering away from other non-failing samples.
- The remaining failing and non-failing patients did not change after review and were assigned to the **"abnormal or heart failure"** or **"within normal limits"** classes, respectively.




*Source: Jeffrey J. Nirschl et al. Consensus clustering using WND-CHARM feature vector exhibits three clusters. The consensus clustering dendrogram and class results are shown above the clustergram.*

# Discussion

- The features used by the CNN for classification **aren't immediately interpretable**
- However, a benefit of representation learning approaches is that **they may reveal novel image features**, learned by the CNN, that are relevant to myocardial disease.
- The performance difference likely reflects the contribution of the features learned by the CNN, which are not present in the set of engineered features.
- . Unlike cancer, where the definitive diagnosis is based on tissue pathology and genetic or molecular markers, **heart failure is a clinical syndrome**. In the clinical setting, pathologists are not called upon to determine whether a patient has heart failure given cardiac histopathology.
- Hence, it's reasonably surprising to note that **CNN outperformed pathologists** at detecting clinical heart failure by a significant margin, **up to 20% in terms of sensitivity and specificity**
- The CNN identified tissue pathology in patients without pre-existing heart failure, suggesting these patients (abnormal cases) may represent cases of **occult cardiomyopathy**, one step closer to predicting the future occurrence of heart failure.

# Limitation and Conclusion

- The classifier was trained on the **extremes of heart disease**: severe heart failure cases requiring advanced therapies (e.g. cardiac transplant or mechanical circulatory devices) versus patients without a history of clinical heart failure. One might argue that **comparing extremes exaggerates classifier performance**.
- However, the **identification** of tissue pathology **in a small subset of patients** without a definitive clinical diagnosis suggests these algorithms are **very sensitive to pathological features of myocardial disease,** and may aid in the detection of disease prior to definitive clinical diagnosis
- A CNN classifier can detect heart failure and show that cardiac histopathology is sufficient to identify patients with clinical heart failure accurately.
- Future work can be focused on predictive modeling indicating digital histopathology with disease progression, survival and treatment responses

# Pancreas Segmentation in MRI Using Graph-Based Decision Fusion on Convolutional Neural Networks (2016)

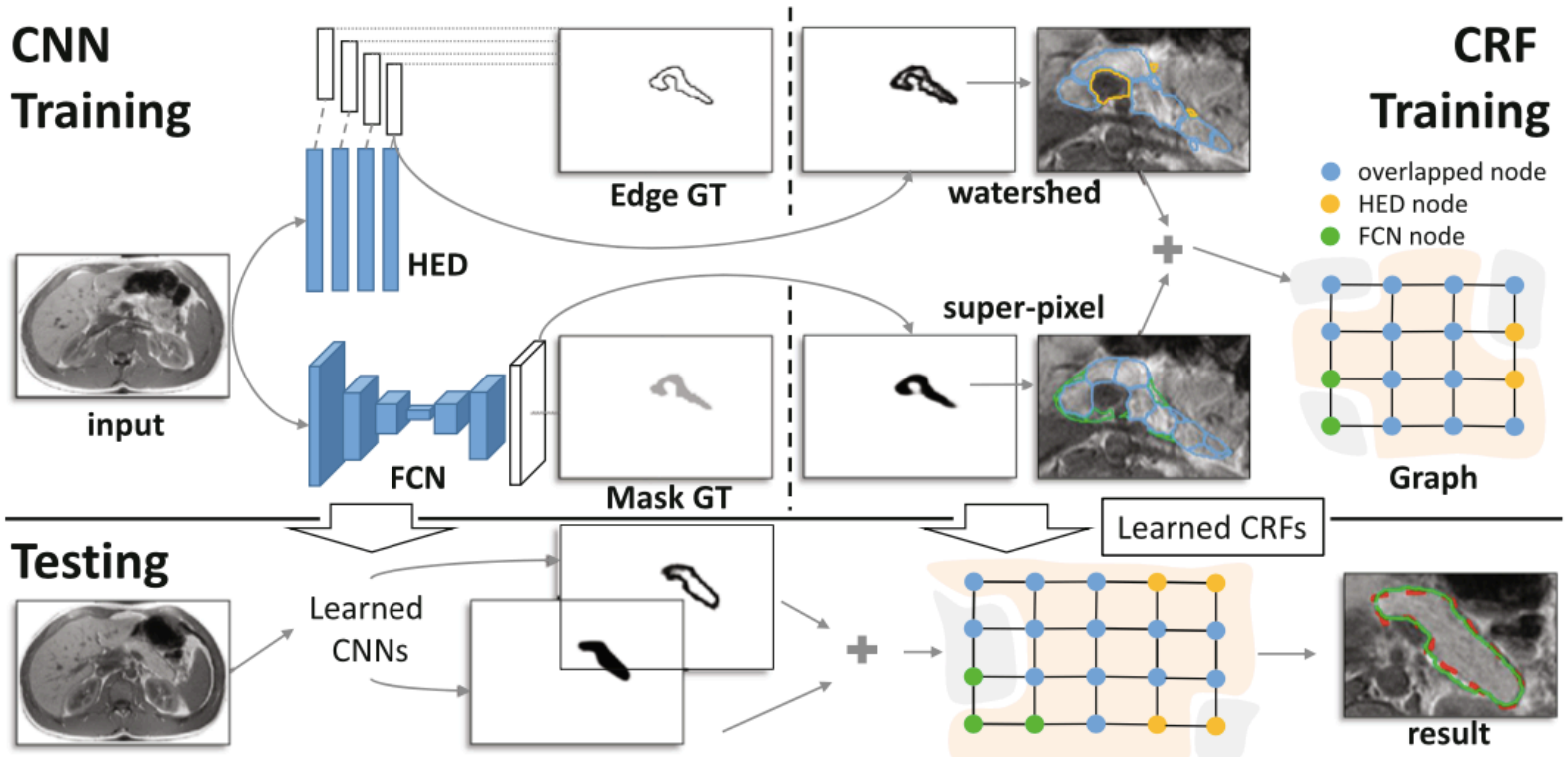
- JINZHENG CAI, LE LU, ZIZHAO ZHANG, FUYONG XING, LIN YANG, QIAN YIN

# Introduction

- Automated segmentation of pancreas remains a challenging problem due to the following:
  - **large appearance variations** in both shape and size of pancreas
  - **highly deformable** as it is relatively soft, can get pushed by its surrounding organs
  - **collapsible boundaries** which causes ambiguities along the boundaries
- Proposed Model: A graph based decision fusion appproach combined with deep CNNs for pancreas segmentation in MRI scans
- Pancreatic detection and boundary segmentation using two types of CNN models:
  - **Pancreatic tissue allocation** with spatial intensity context
  - **Boundary detection** to allocate the semantic boundaries of pancreas
- The CNN results then combined with a **graph based decision fusion model** to obtain the refined segmentation outputs

# Pipeline

**CNN Training**: CNN models are trained for pancreatic tissue allocation (FCN model) and boundary detection (the HED model);
**CRF Training**: A CRF model is learned based on the candidate regions that are detected by CNN models.
**Testing**: The segmentation begin with CNN models, and further refined by the CRF model. It followed iterated conditional modes algorithm to perform the graph inference.




**Note**: ICM is a deterministic algorithm for obtaining a configuration of a local maximum of the joint probability in a CRF

*Source: Jinzheng Cai et al.*

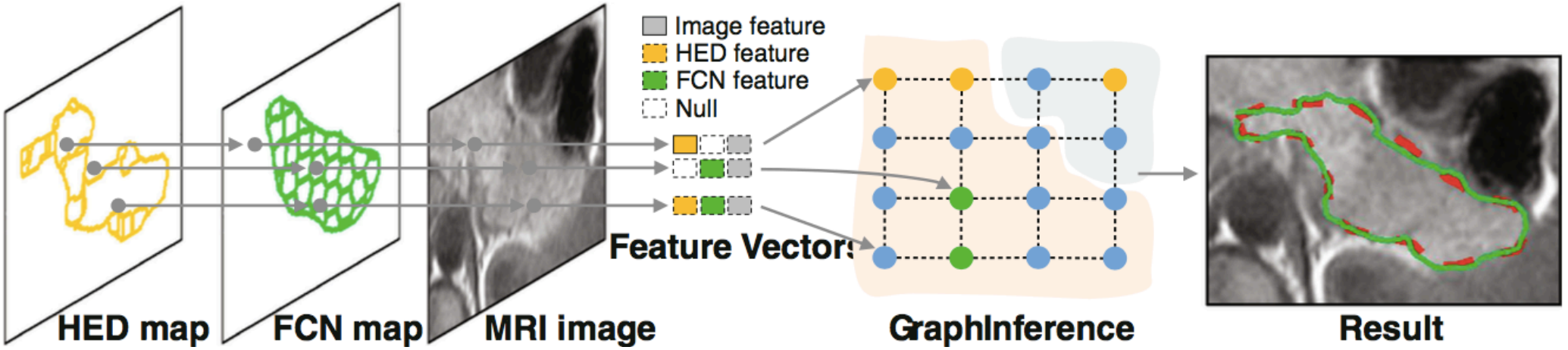
# Design of CNNs

- Fine tuning both the CNNs from the **pre-trained VGG-16 network**
- **Pixel level classification** strategy too computationally **expensive**
- Fully Convolutional Network (FCN) with an end to end scheme, consisting of the **upsampling layers**, which increase the output resolution to the level of the input image
- For a robust object level boundary detection, **holistically-nested edge detection** (HED) is used
- HED improves a regular FCN by adding **deep supervision at all convolutional layers** against down-sampled maps of the final desirable labeling output
- Deep Supervision: Initializing the **parameters** of the network with those of **pre-trained shallower networks** or **add auxiliary classifiers** connected to intermediate layers

# Decision Fusion

Two components:

- Graphical Model: overview in the figure given below
- Conditional Random Field




*Source: Jinzheng Cai et al.*

- HED map: **Watershed transformation** of the HED output (semantic gradient map)
- FCN map: **Superpixel segmentation** on the detected FCN regions

# Decision Fusion: Graphical Model

- In reference to the overview image above:
- Undirected graph with weighted edges
- Feature vector of each node extracted from the corresponding region of the HED map, the FCN map, and the MRI image in order (null: non detected region from HED/FCN maps)
- Node feature extraction to begin with the HED map so as to preserve as many detected boundaries as possible
- Adjacent nodes linked with weighted edges: reflecting the likelihood of the two nodes belonging to the same category (pancreas or non-pancreas)
- Edges with low weight (similarity): nodes of different category
- Nodes with both non-null features: higher likelihood of being the pancreas region

# Decision Fusion: Conditional Random Field

- Classification likelihood depends on: FCN and HED map features & handcrafted features of the MRI image
- CRFs: Probabilistic framework for labeling & segmenting data
- CRF objective function:
- $\boldsymbol{v} = [v_1, ..., v_N]$: vector containing labels for all the nodes; $v_i \in \{0, 1\}$: states (pancreas or non-pancreas)
- $u_{ij} = \mathbf{1}[v_i = v_j]$; $(\alpha_0, ..., \alpha_K, \beta_0, ..., \beta_K)$ are parameters of CRF model to be learned (via Stochastic Gradient Descent)
- $\boldsymbol{f_i}$ is the feature vector of the i-th node
- $S_{[FCN,HED,i]}$ denotes the area of FCN map, HED map or i-th node
- $I_i$ : mean intensity value; and $h_i$ : histogram from the pixels of that region

$$E(\boldsymbol{v}) = \sum_i \phi_u(v_i) + \sum_{(i,j)\in\mathcal{N}} \phi_p(v_i, v_j)$$

$$\phi_u(v_i) = exp(\alpha_0 + \sum_{k=1}^{K} v_i \alpha_k f_{ik}),$$

$$\phi_p(v_i, v_j) = exp(u_{ij}(\beta_0 + \sum_{k=1}^{K} \frac{\beta_k}{1 + ||f_{ik} - f_{jk}||_2}))$$

$$\boldsymbol{f_i} = \left[\frac{|S_{FCN} \cap S_i|}{|S_i|}, \frac{|S_{HED} \cap S_i|}{|S_i|}, I_i, h_i\right]$$

# Experimental Details

- Abdominal MRI scans captured from 78 subjects; 52: training, 26: testing
- For each scan, manual annotation of the pancreas was given by a board-certified radiologist
- Dice Similarity Coefficient was used compare the similarity of manual annotation ($L_1$) versus automated segmentation ($L_2$) results

$$DSC = 2|L_1 \cap L_2|/(|L_1| + |L_2|)$$

- Fine tuning CNN models:
  - FCN: initial learning rate: $1.0 \times 10^{-2}$, weight decay: 0.1 for every $5.0 \times 10^4$ iterations, max. iteration: $1.5 \times 10^5$
  - HED: initial learning rate: $1.0 \times 10^{-6}$, weight decay: 0.1 for every $5.0 \times 10^4$ iterations, max. iteration: $4.0 \times 10^4$
  - Output as a probability likelihood map where each location beyond a threshold belongs to pancreas

# Experimental Details

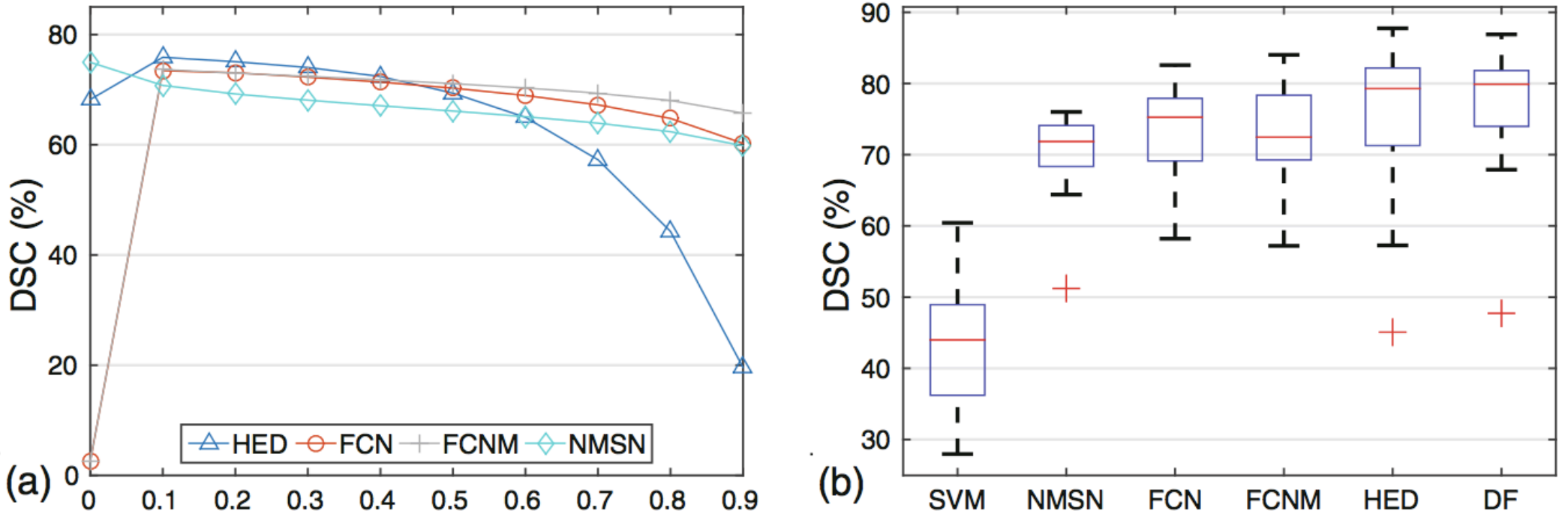
- Segmentation baseline on the MRI dataset: Extracted **HoG** features from 64 × 64 image patches, and conducted **superpixel-wise prediction with SVM**
- CNN trained from scratch: **11-layer** (adjoining table) Neuronal Membranes Segmentation Network (NMSN) model made **pixel-wise prediction** on 95×95 image patches

| Layer | Type | Maps and neurons | Kernel size |
|---|---|---|---|
| 0 | input | 1 map of 95x95 neurons | |
| 1 | convolutional | 48 maps of 92x92 neurons | 4x4 |
| 2 | max pooling | 48 maps of 46x46 neurons | 2x2 |
| 3 | convolutional | 48 maps of 42x42 neurons | 5x5 |
| 4 | max pooling | 48 maps of 21x21 neurons | 2x2 |
| 5 | convolutional | 48 maps of 18x18 neurons | 4x4 |
| 6 | max pooling | 48 maps of 9x9 neurons | 2x2 |
| 7 | convolutional | 48 maps of 6x6 neurons | 4x4 |
| 8 | max pooling | 48 maps of 3x3 neurons | 2x2 |
| 9 | fully connected | 200 neurons | 1x1 |
| 10 | fully connected | 2 neurons | 1x1 |

*Source: Dan C. Ciresan et al.*

- To compare with deep CNN models, an FCN model (FCNM) was trained with **3 classes** of pancreatic tissue, pancreatic boundary and the background
- To generate the graphical model for Decision Fusion (DF), graph nodes **with more than 50% overlaps** with the human annotations were assigned positive, rest negative

# Results




*Source: Jinzheng Cai et al.*

a) Mean DSC w.r.t. thresholds on the output probability (plateau region: [0.2, 0.6])
b) Comparison of segmentation accuracy; red lines: means, crosses: outliers

# Results

- Decision Fusion (DF) approach achieves the highest accuracy w.r.t. mean DSC
- Second minimal standard deviation of the DF approach
- NMSN, trained from scratch, has the smallest variance; though shallow architecture renders poor performance
- This approach reported the **best quantitative pancreas segmentation performance** with a mean DSC 76.1% ($\pm$ 8.7) as of 2016
- Although other past results cannot be strictly campared due to the lack of any common evaluation datasets

# References


- Geert Litjens et al.: A survey on deep learning in medical image analysis (2017)
- Jeffrey J. Nirschl et al.: A deep-learning classifier identifies patients with clinical heart failure using whole-slide images of H&E tissue (2018)
- Nikita Orlov et al.: WND-CHARM: Multi-purpose image classification using compound image transforms (2008)
- Jinzheng Cai et al.: Pancreas Segmentation in MRI Using Graph-Based Decision Fusion on Convolutional Neural Networks (2016)
- S.V.N. Vishwanathan et al.: Accelerated training of conditional random fields with stochastic gradient methods (2006)
- S. Xie et al.: Holistically-nested edge detection
- Dan C. Ciresan et al.: Deep neural net- works segment neuronal membranes in electron microscopy images

# Thank You.