# Challenges in Applying Deep Learning Methods for Medical Image Processing

Fatemeh Zabihollahy, Ph.D.
Postdoctoral Research Fellow
University of California, Los Angeles

# Common Challenge in Applying DL-based Methods for Medical Image Analysis

1) Limited number of annotated images for training

2) Class imbalance

3) Performance often degrades when algorithms are applied on new data acquired from different scanners or sequences than the training data

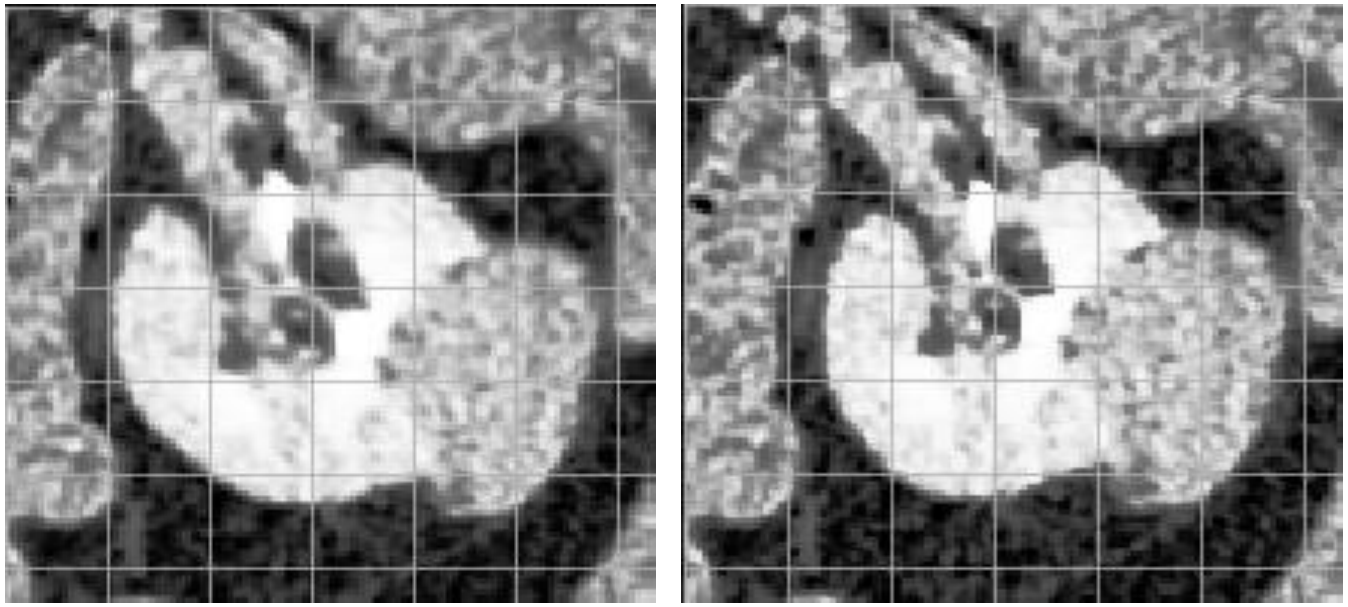
# Data Augmentation for Medical Images

- Deep neural networks (DNN) typically have parameters in the order of millions.
- Naturally, if we have a lot of parameters, a proportional number of examples are needed, to get good performance.
- Recalling that your neural network is only as good as the data you feed it.

# Popular Augmentation Techniques

- Flip (up-down, left-right)
- Rotate
- Crop
- Shift
- Scale
- Gaussian Noise
- Conditional GANs
- Elastic Distortion (Simard et al., ICDAR 2003)

# Example of Elastic Distortion



From left to right, an example of one training sample with benign renal mass and its deformed one are shown.

Note: for image segmentation, image and ground truth must be augmented similarly.

# Data Augmentation Pipelines [A. Gandhi, 2020, https://nanonets.com/]

- Offline:

➢ It performs all necessary transformations beforehand, to increase the size of the dataset.

➢ This method is preferred for relatively smaller datasets.

➢ The size of the dataset is increased by a factor equal to the number of transformations are performed.

- Online (augmentation on the fly):

➢ It performs all transformations on a mini-batch, just before feeding it to the machine learning model.

➢ This method is preferred for larger datasets, as the explosive increase in size cannot be afforded.

# Transfer Learning [Jain, educative.io]

- Human have an inherent ability to transfer knowledge across tasks.
- We know how to ride a bike → learn how to drive a car.
- Transfer learning is the idea of utilizing knowledge acquired for one task to solve related ones.
- In transfer learning, we can leverage knowledge (features, weights, etc.) from previously trained models for training new models to address the problem of having few labeled data for the newer task.
- Transfer learning works because certain low-level features such as edges, shapes, and corners can be shared across tasks.

# Transfer Learning Cont.

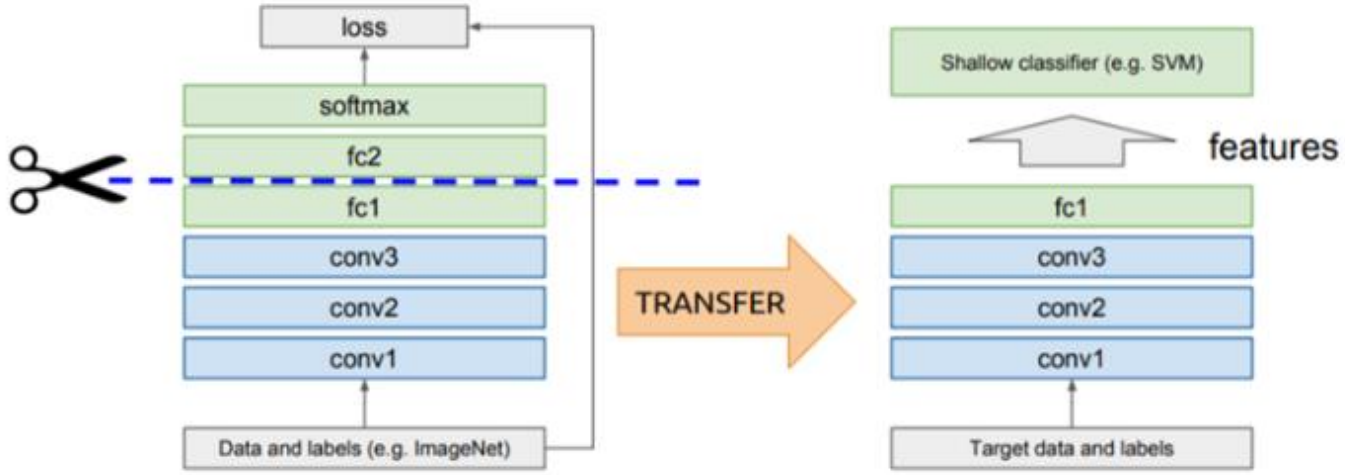
During the process of transfer learning, the following questions must be answered:

- **What to transfer**: what part of the knowledge is source-specific and which part is common between the source and target.
- **How to transfer**: which strategy to choose to transfer knowledge across domains/tasks.
- **When to transfer**: there are scenario where transferring knowledge does not improve the learning but makes it worst.

# Transfer Learning Strategies

**1) Pre-trained model as feature extractors**

- Remove the last layer
- The weights are not getting update




**2) Fine tuning off-the-shelf pre-trained models**

- Selectively retrain some of the previous layers
- Freeze certain layers while retraining the rest

**Retrieved from: towardsdatascience.com, A Comprehensive Hands-on Guide to Transfer Learning with Real-World Applications in Deep Learning**

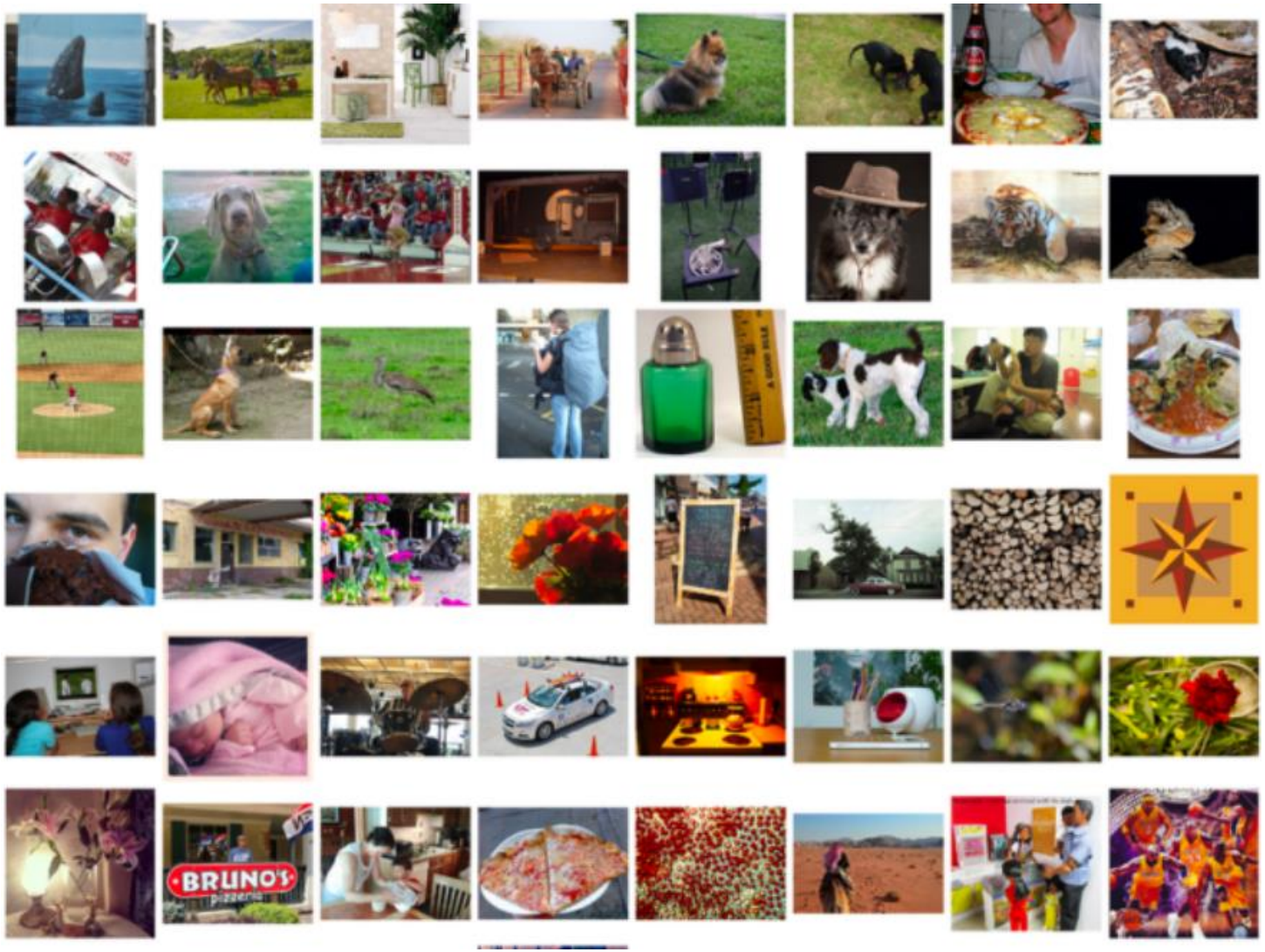
**3) Two-stage transfer learning**

- Newly attached fully layer is trained while freezing the rest layer for few epochs
- Some other layers are unfrozen and retrained

# Models for Transfer Learning

You can leverage some popular models including VGG (e.g., VGG16 or VGG19), GoogLeNet (e.g., InceptionV3), Residual Network (e.g., ResNet50), MobileNet

For the classification task, images must be classified into one of 1,000 different categories.




Sample of ImageNet

# Limitation of Transfer Learning for Medical Images

[Raghu t al., “Transfusion: Understanding Transfer Learning for Medical Imaging”, 2019]

“Transfer learning will not work well if the distribution of the images in our task is drastically different from the images that the base network was trained on.”

“There are fundamental differences in data sizes, features and task specifications between natural image classification and the target medical tasks.”

“Having benchmarked both standard ImageNet architectures and non-standard lightweight models on two large scale medical tasks, the authors found that transfer learning offers limited performance gains and much smaller architectures can perform comparably to the standard ImageNet models.”

# Neural Architecture Search (NAS)

- Recently, Neural Architecture Search (NAS) was proposed to search for the best network architecture for a given problem, automatically.

- However, the inconsistency between search stage and deployment stage often exists in NAS algorithms due to memory constraints and large search space.

# NAS Cont.

- NAS can be seen as a subfield of AutoML and has a significant overlap with hyperparameter optimization.
- The objective of NAS is typically to find architectures that achieve high predictive performance on unseen data.

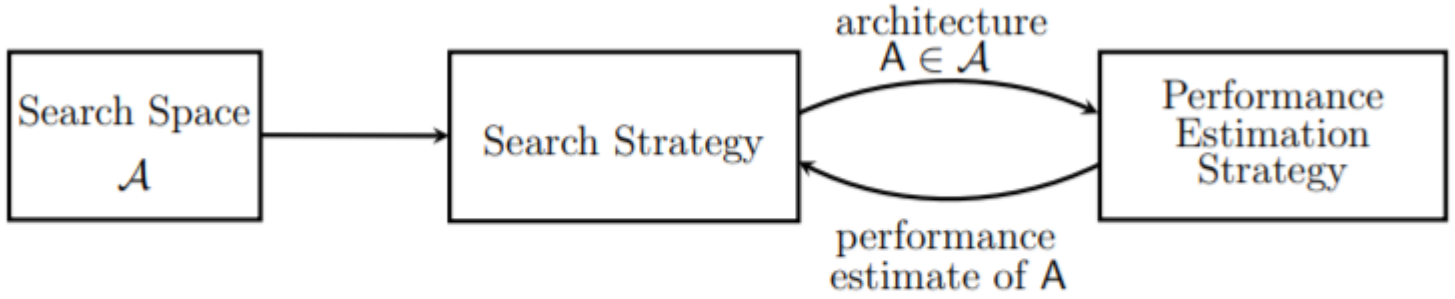



Figure 1: Abstract illustration of Neural Architecture Search methods. A search strategy selects an architecture $\mathsf{A}$ from a predefined search space $\mathcal{A}$. The architecture is passed to a performance estimation strategy, which returns the estimated performance of $\mathsf{A}$ to the search strategy.

[Elsken et al., Journal of Machine Learning Research 2019]

# Search Space

- Defines which architectures can be represented in principle.
- Incorporating prior knowledge can reduce the size of the search space and simplify the search.
- However, this also introduces a human bias, which may prevent finding novel architectural building blocks that go beyond the current human knowledge.

# Search Strategy

- Details how to explore the search space.
- Encompasses the classical exploration-exploitation trade-off since, on the one hand, it is desirable to find well-performing architectures quickly, while on the other hand, premature convergence to a region of suboptimal architectures should be avoided.
- The exploration-exploitation trade-off that is a well-known problem in reinforcement learning and recommendation systems occurs in scenarios where a learning system has to repeatedly make a choice with uncertain pay-offs.
- The dilemma for a decision-making system with incomplete is whether to repeat decisions that have worked well so far (exploit) or to make novel decisions, hoping to gain even greater rewards (explore).

# Performance Estimation Strategy

- Refers to the process of performing a standard training and validation of the architecture on data, but this is unfortunately computationally expensive and limits the number of architectures that can be explored.

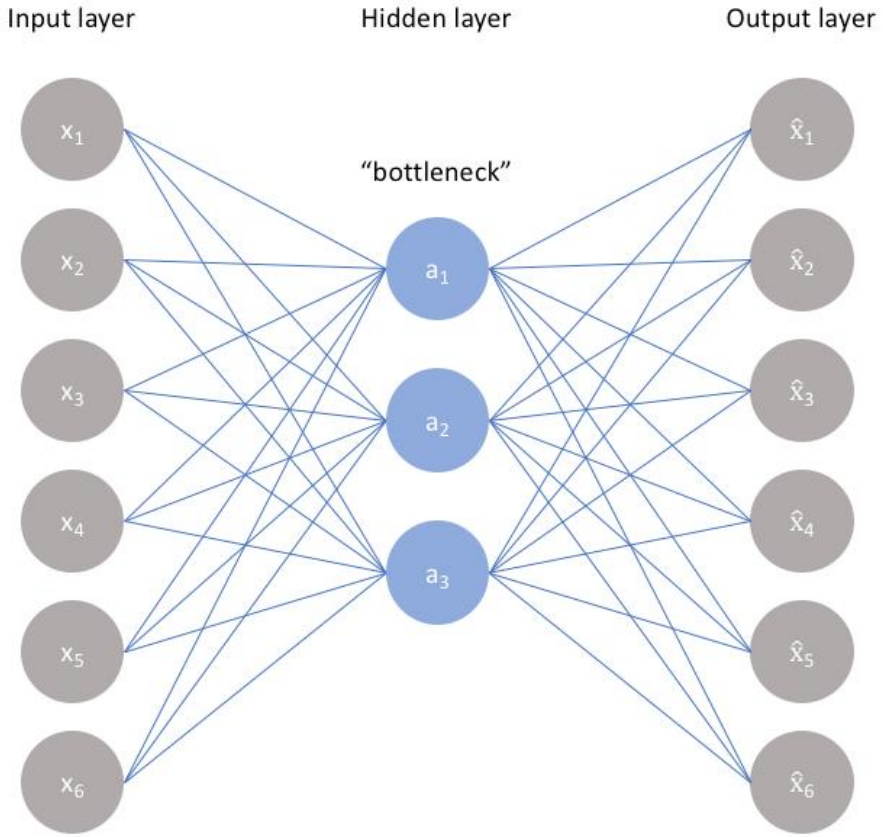
- Much recent research therefore focuses on developing methods that reduce the cost of these performance estimations.

# Self-supervised Learning (SSL)

- Good performance of deep learning models usually requires a decent amount of labeled data but collecting annotated medical images are very expensive.
- On the other hand, the amount of unlabeled data is substantially more than a limited number of manually labelled data.
- It is crucial to learn from unlabeled data; however, unsupervised learning is not easy and usually works much less efficiently than supervised learning
- We can get labels for free for unlabeled data and train unsupervised dataset in a supervised manner, which is referred to as SSL.
- The major techniques in SSL are generative and discriminative models.

# Generative Methods

- These methods aim at the accurate reconstruction of data after passing it through a bottleneck (e.g. autoencoder).
- The input is reduced into a low dimensional space using an encoder network and the image is reconstructed using the decoder network.
- The input itself becomes the supervision signal for training the network.
- The encoder part can be used as a starting point to build a classifier, using on the transfer learning strategies.

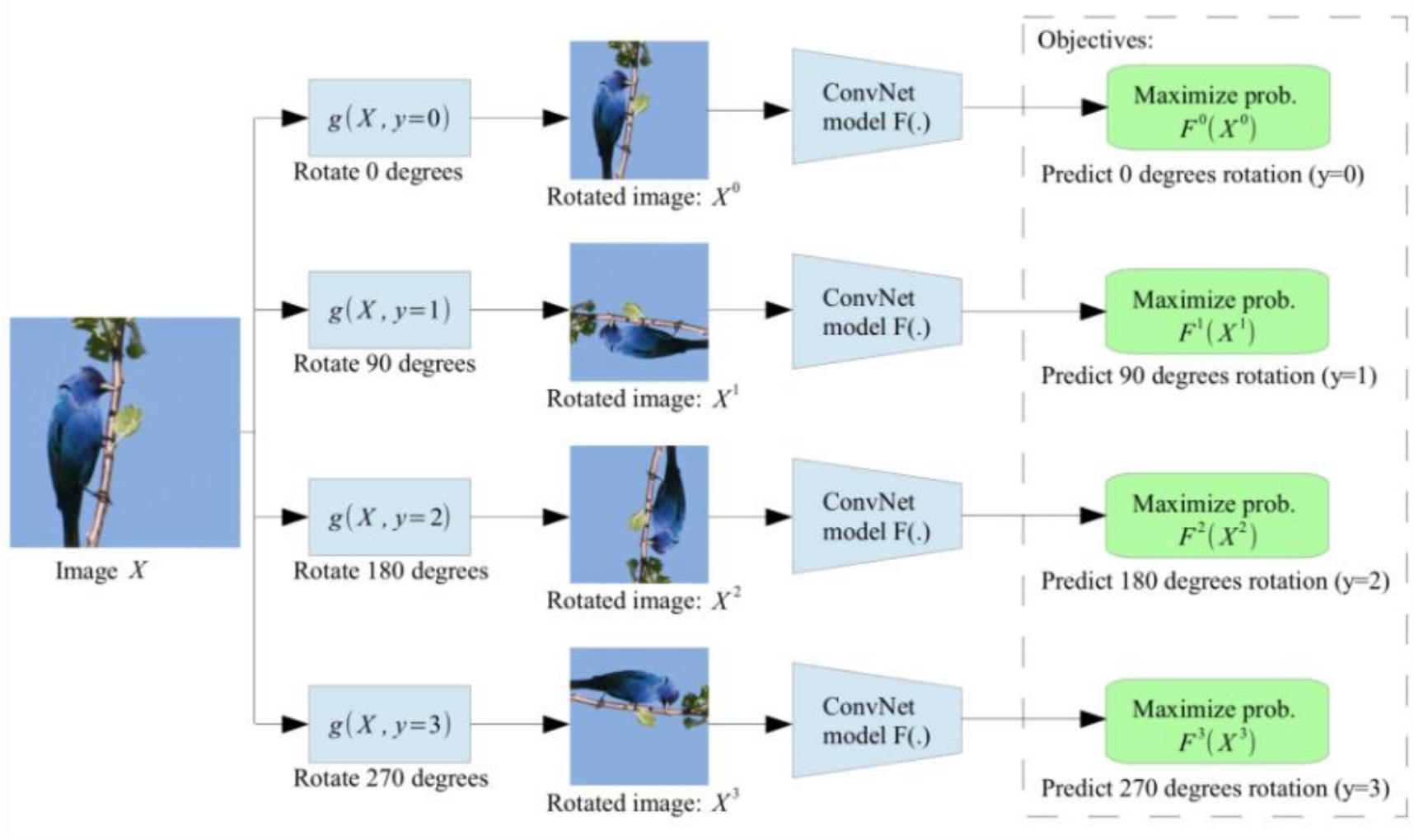



Retrieved from
https://www.jeremyjordan.me/autoencoders/

# Discriminative Methods

- Different labeling that are freely available besides or within visual data are exploited and used as intrinsic reward signals to learn general-purpose features.
- These approaches train a NN to learn an auxiliary classification task.
- An auxiliary task is chosen such that the supervision signal can be derived from the data itself, without human annotation.
- The features obtained with these approaches have been successfully transferred to classification and detections tasks, and their performance is very encouraging when compared to features trained in a supervised manner.

# Learning the geometric transformations applied on image




[Gidaris et al. 2018]

# Learning of Visual Representations by Solving Jigsaw Puzzles

[Noroozi et al., 2017]

# Class Imbalance In Dataset

Training:

- Data Augmentation

Testing:

- Select proper metric to evaluate the performance of the method

# Domain Adaptation [Kouw et al., 2019]

- Recent advances in DL-based methods have come to define the state-of-the-art for many medical imaging applications.

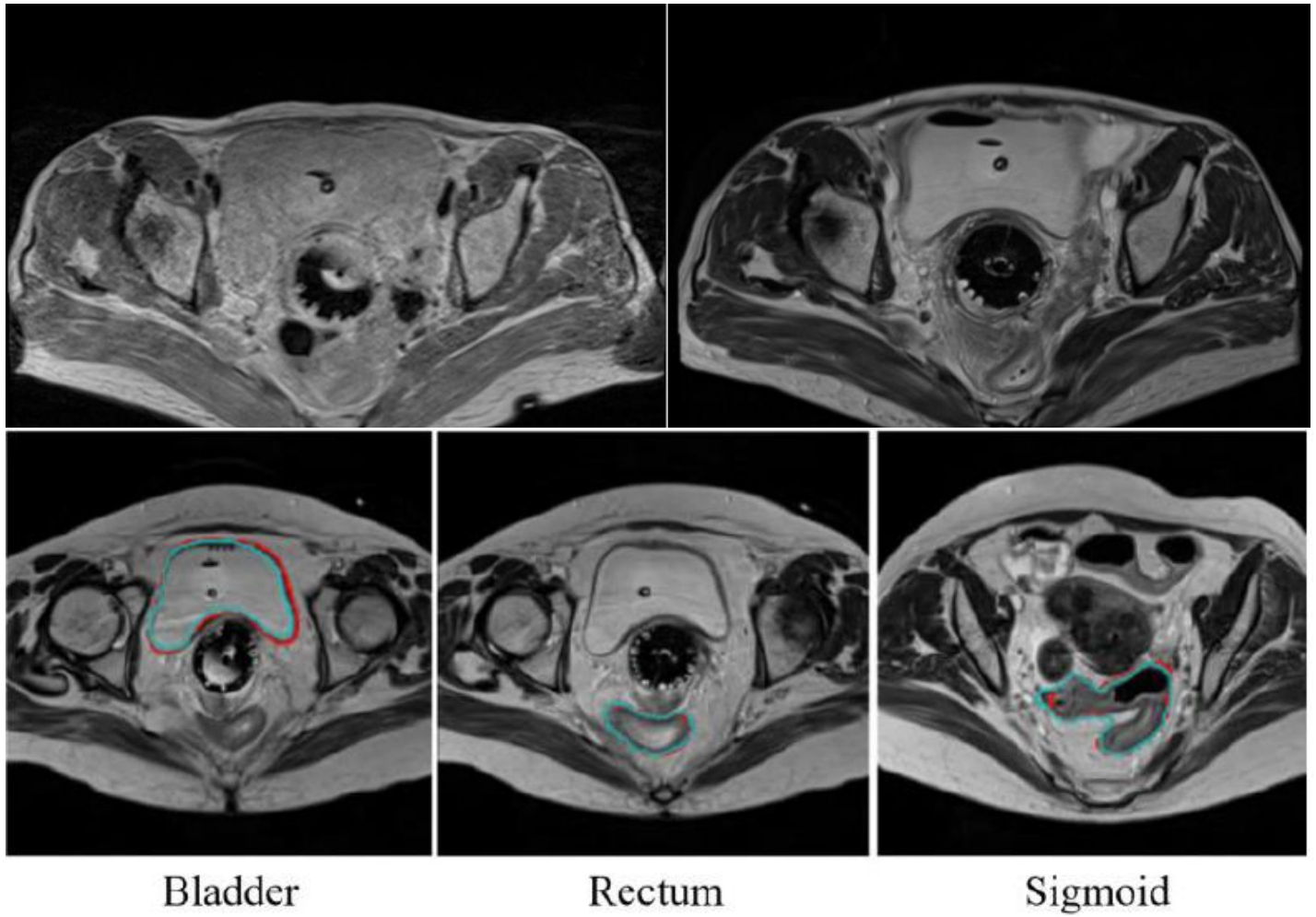
- Those models, however, fail to generalize when applied to other domains, a very common scenario in medical imaging due to the variations in multi-center data in medical imaging studies, the variability of images and anatomical structures, even across the same imaging modality that has brought the necessity of domain adaptation.

- We define domains as the combination of an input space X, an output space Y, and an associated probability distribution p.

- Domain adaptation is defined as the particular case where X and Y remain unchanged and only the probability distributions change.

# Domain Adaptation for Medical Image Segmentation (Supervised)

- Domain adaptation can be categorized into three settings, supervised, semi-supervised, and unsupervised.
- In supervised domain adaptation methods, labeled data from the target domain is used.
- In the context of convolutional neural networks (CNNs), supervised domain adaptation can be approached by training from scratch or fine-tuning a network pre-trained on the source domain.




[Zabihollahy et al., 2021]

# Regularization

- Regularization is the process of adding information in order to prevent overfitting in ML problems.
- The foremost problem in ML is to develop an algorithm that not only performs well on training data but also on the new data.
- Different strategies help to reduce generalization error.
- These strategies are known collectively as regularization.
- Major forms of regularization that are available for DL practitioners are dataset augmentation, parameter sharing, dropout, early stopping, batch normalization, and ensemble methods.

# Data Augmentation

- The common approach for reducing generalization error in ML is to train the model on more data.
- The problem is that in practice the number of training instances is limited.
- To tackle this, the artificial data can be created from the original data and augmented to the training set.
- For image data, artificial images can be generated through different transformations such as image shifting, rotating, and flipping.
- Artificially augmenting training data is also beneficial for addressing class imbalance issue.

# Parameter Sharing

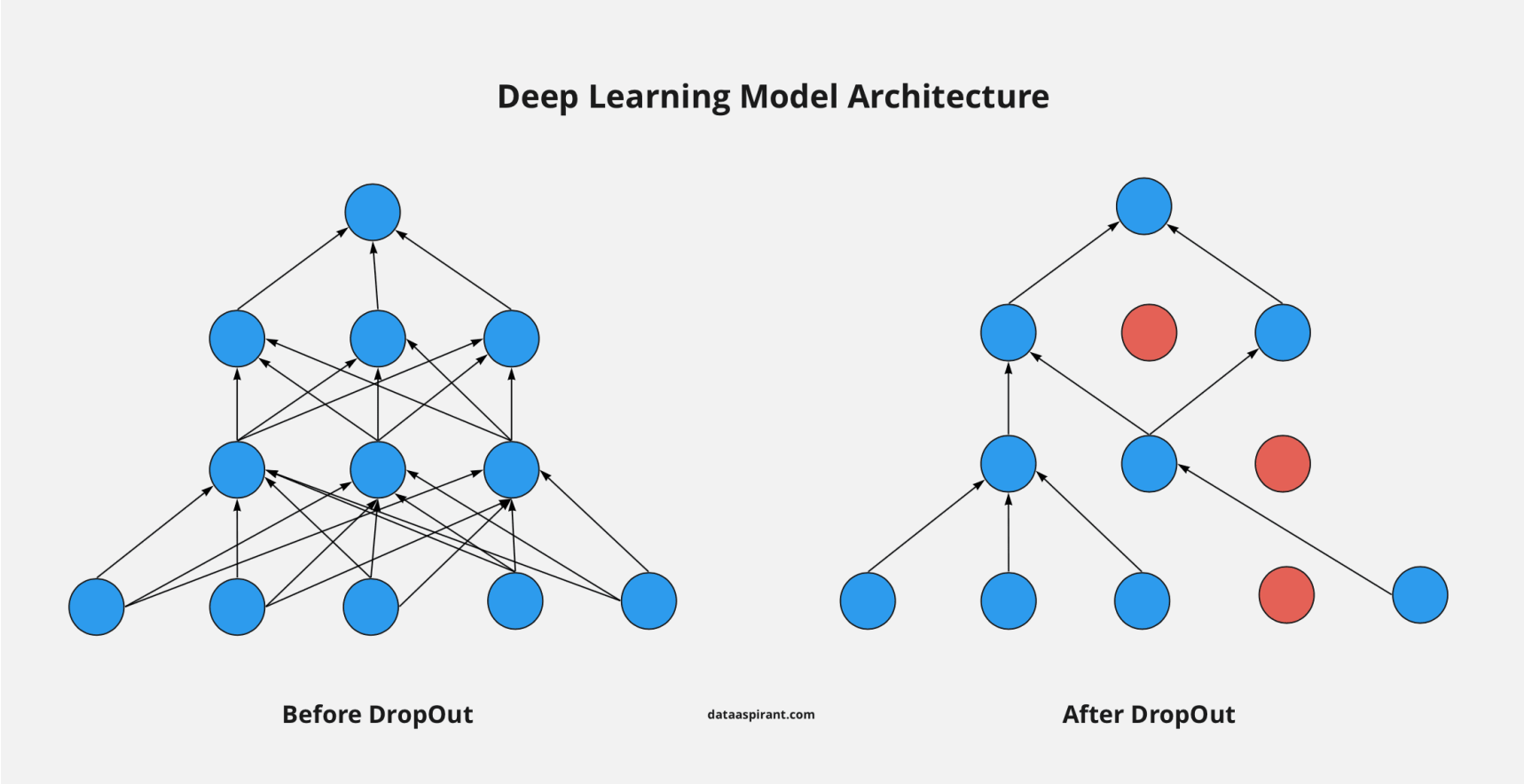
- By nature, an extensive parameter sharing occurs in CNN as each kernel in the convolutional layer is moved across the whole input data at every position.

- Thus, in CNN only one set of kernel is learned rather than learning a separate set of parameters for every location.

- Parameter sharing not only regularizes the parameters but also leads to a significant reduction in the required memory by storing only a subset of the parameters.

# Dropout

- Randomly drop some nodes along with their connections in the training phase, so each iteration has a different set of nodes.
- During the testing phase, the average of all those networks' predictions is obtained.




Retrieved from https://dataaspirant.com/8-deep-learning-dropout/

# Early Stopping

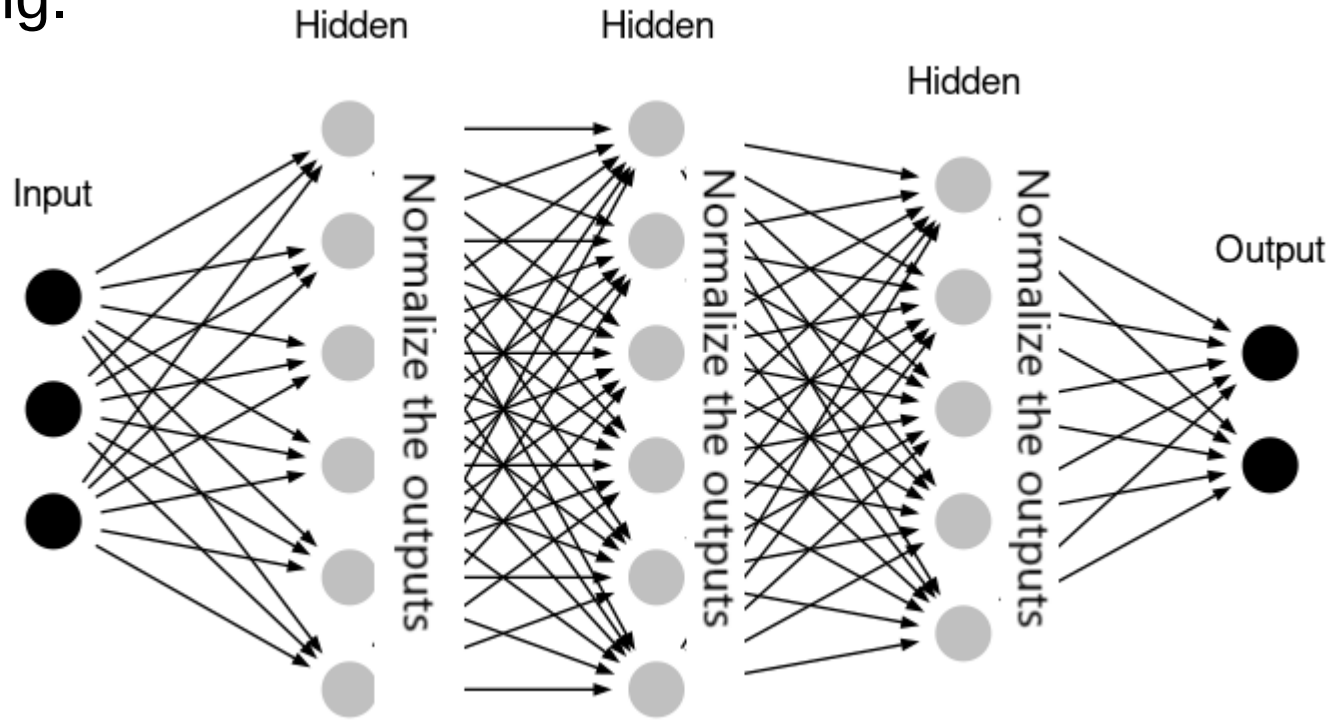
- Investigating the train and validation errors during training a model shows that after a certain number of epochs, the validation error begins to increase while the training error steadily decreases.

- To obtain the best-trained model, we can store the model parameter with the lowest validation set error and use it for the test rather than the latest parameters, aiming to get a better result on the test set.

- This strategy is known as early stopping, which is one of the most common regularization techniques in DL due to its simplicity and efficiency.

# Batch Normalization

- Batch normalization is another regularization method that keeps the mean and standard deviation of the input data close to zero and one, respectively.
- This operation makes the learned function invariant to scaling of the weights.
- Batch normalization makes the optimization significantly smoother, which in turn leads to more predictive and stable behavior of the gradient, allows for faster training.




Retrieved from https://deepai.org/machine-learning-glossary-and-terms/batch-normalization

# Ensemble Learning

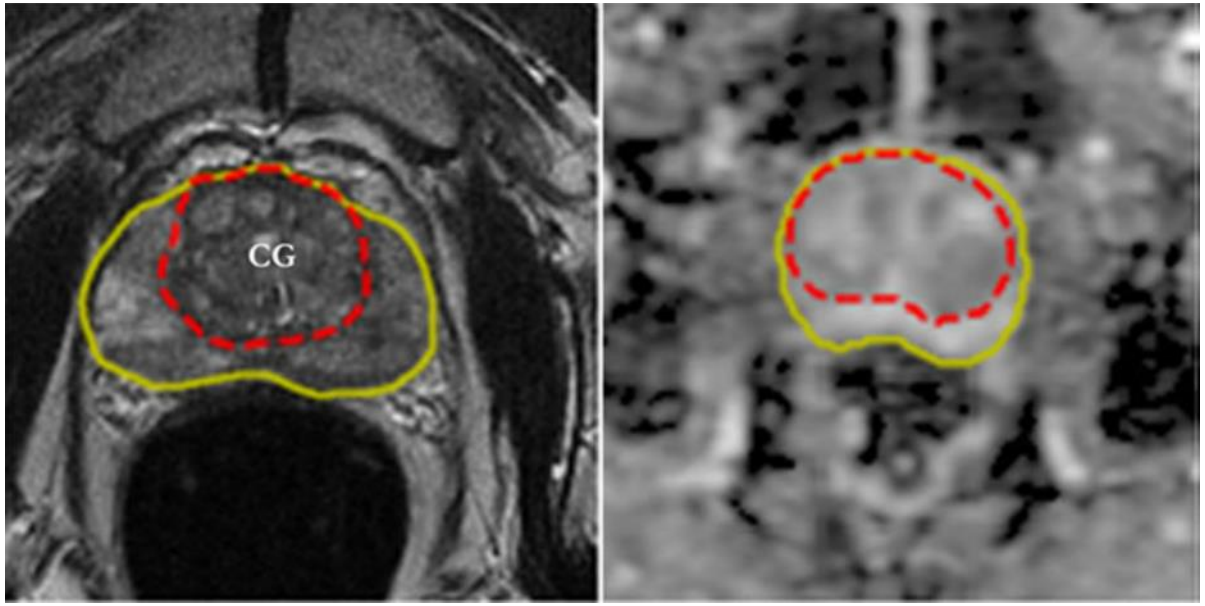
- In ensemble learning, multiple classifiers of the same or different kinds are trained and combined for classification task to obtain a better predictive performance or reduce the chance of selecting a poor model for a given classification task.

- The most popular algorithms in ensemble learning are Bagging, Boosting, Adaboost, and a mixture of experts in which the outputs of several classifiers are combined through a linear rule.

- This rule could be an algebraic combiner such as minimum, maximum, sum, average, etc. or a voting-based combiner, where majority voting or weighted majority voting is employed for final prediction.

# Example of my work on ensemble Learning

- Prostate cancer (PCa) is the most common non-cutaneous malignancy in men [1].
- Prostate MRI has been shown to be accurate for diagnosis of PCa, particularly in the PZ [2].
- Generally, developed CAD tools for PCa detection require manual prostate zonal segmentation.



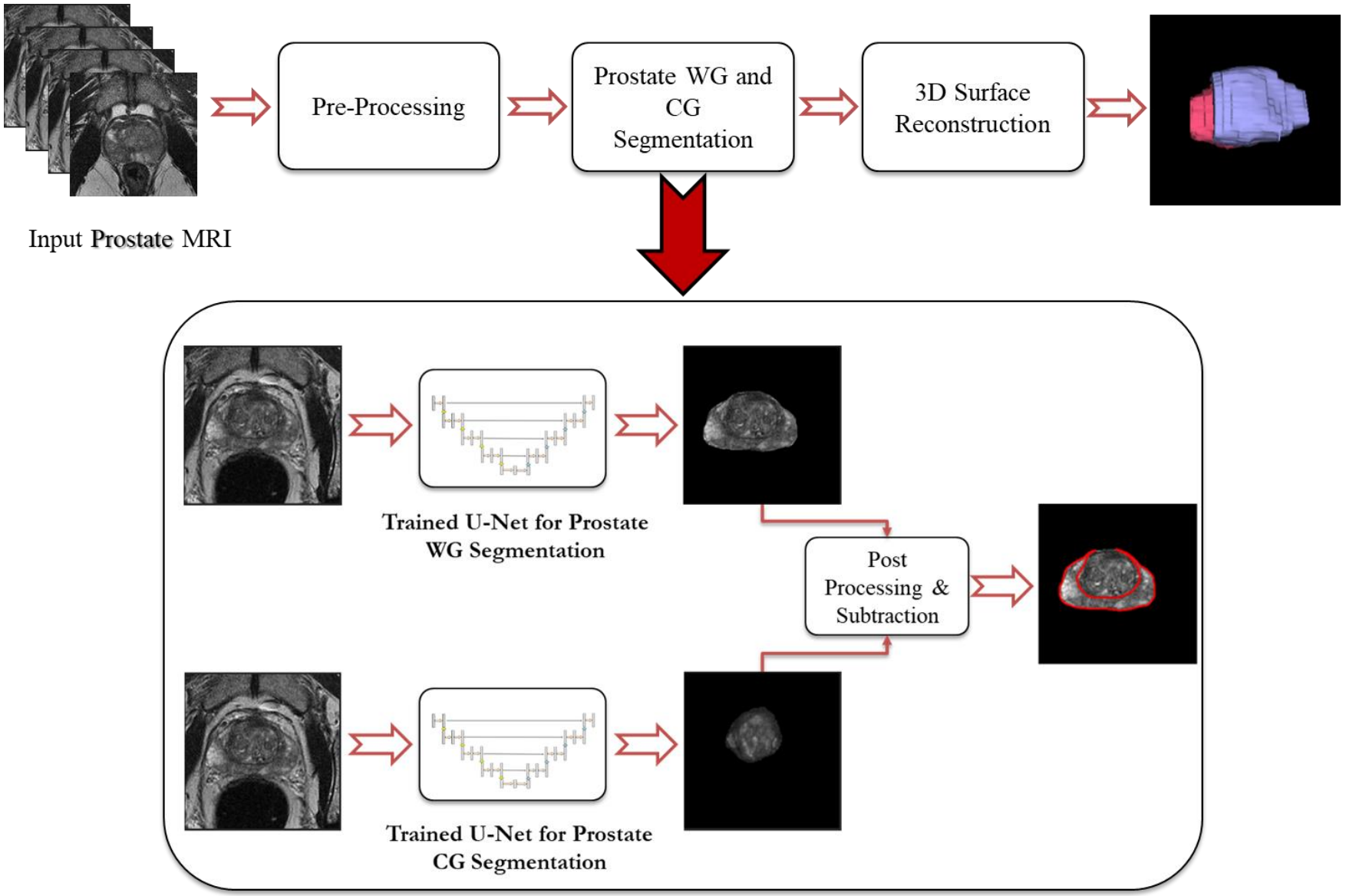

[1] Mistry K, Cable G. Meta-analysis of prostate-specific antigen and digital rectal examination as screening tests for prostate carcinoma. J Am Board Fam Med. 2003;16:95–101.
[2] Eichler K, Hempel S, Wilby J, Myers L, Bachmann LM, Kleijnen J. Diagnostic value of systematic biopsy methods in the investigation of prostate cancer: a systematic review. J Urol. 2006;175:1605–1612.

# Dataset and Experiment Design

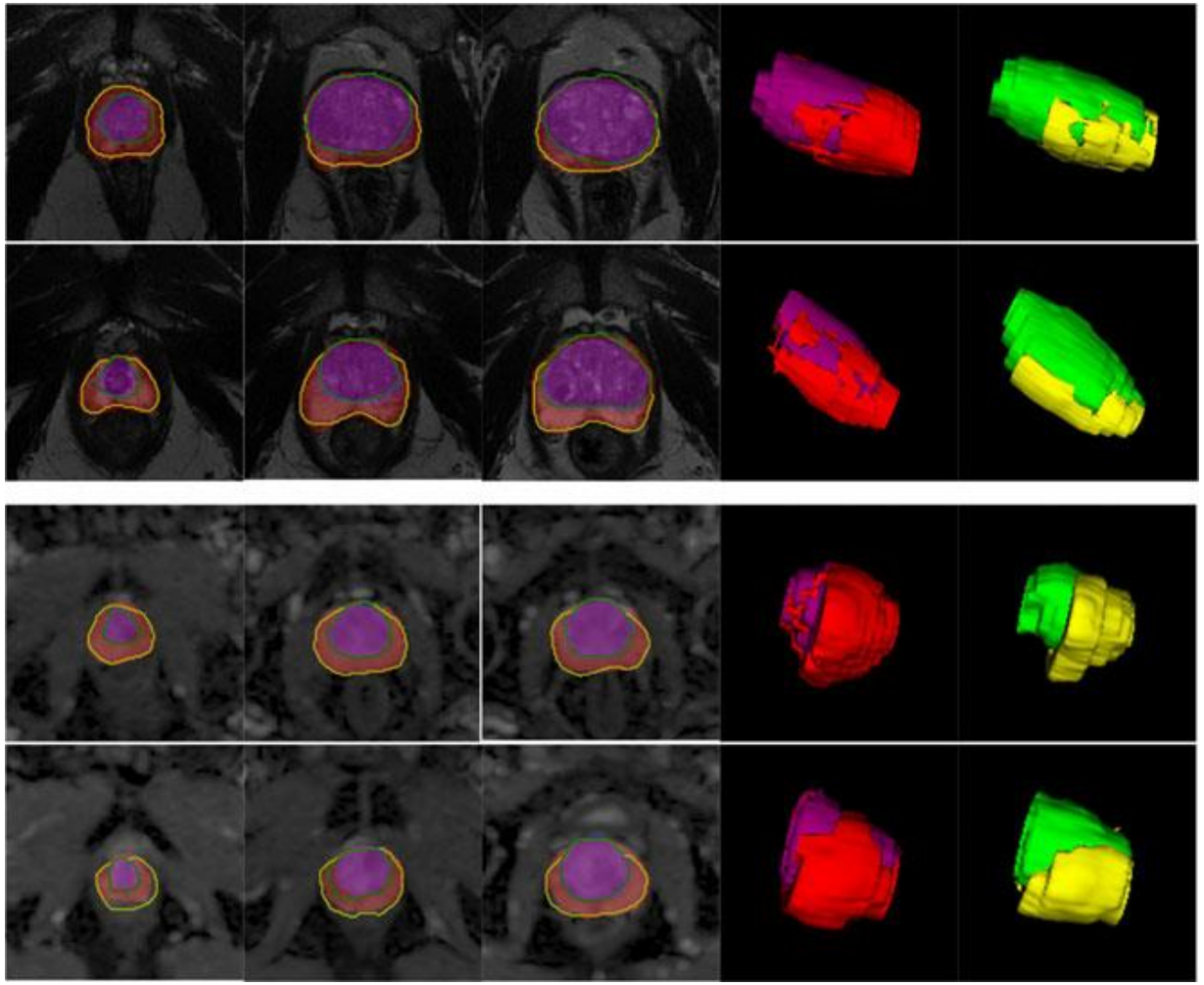
Images were evaluated by a team of four fellowship-trained abdominal and pelvic Radiologists with between 5 and 13 years of experience in prostate mp-MRI.

| MR Image Type | Total number of patients in the dataset | Train | | | Test | | |
|---|---|---|---|---|---|---|---|
| | | Number of patients | Ratio of patients with/without significant tumor | Total No. of compiled 2D slices | Number of patients | Ratio of patients with/without significant tumor | Total No. of compiled 2D slices |
| T2W | 225 | 100 | 70/30 | 1154 | 125 | 83/42 | 1587 |
| ADC map | 225 | 100 | 70/30 | 812 | 125 | 83/42 | 917 |

# The Block Diagram of our Proposed Method for Prostate Regional Segmentation

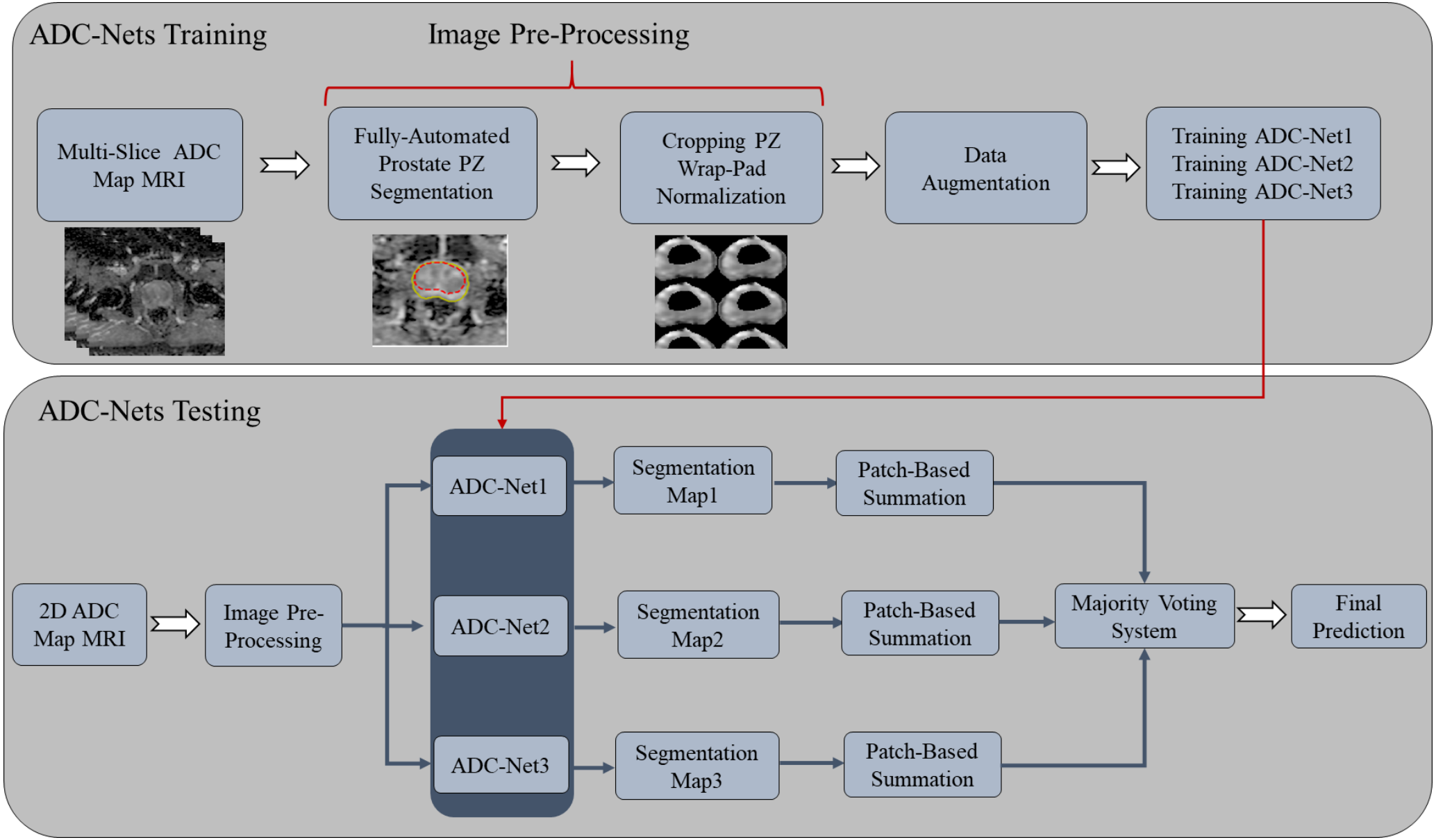


# Results

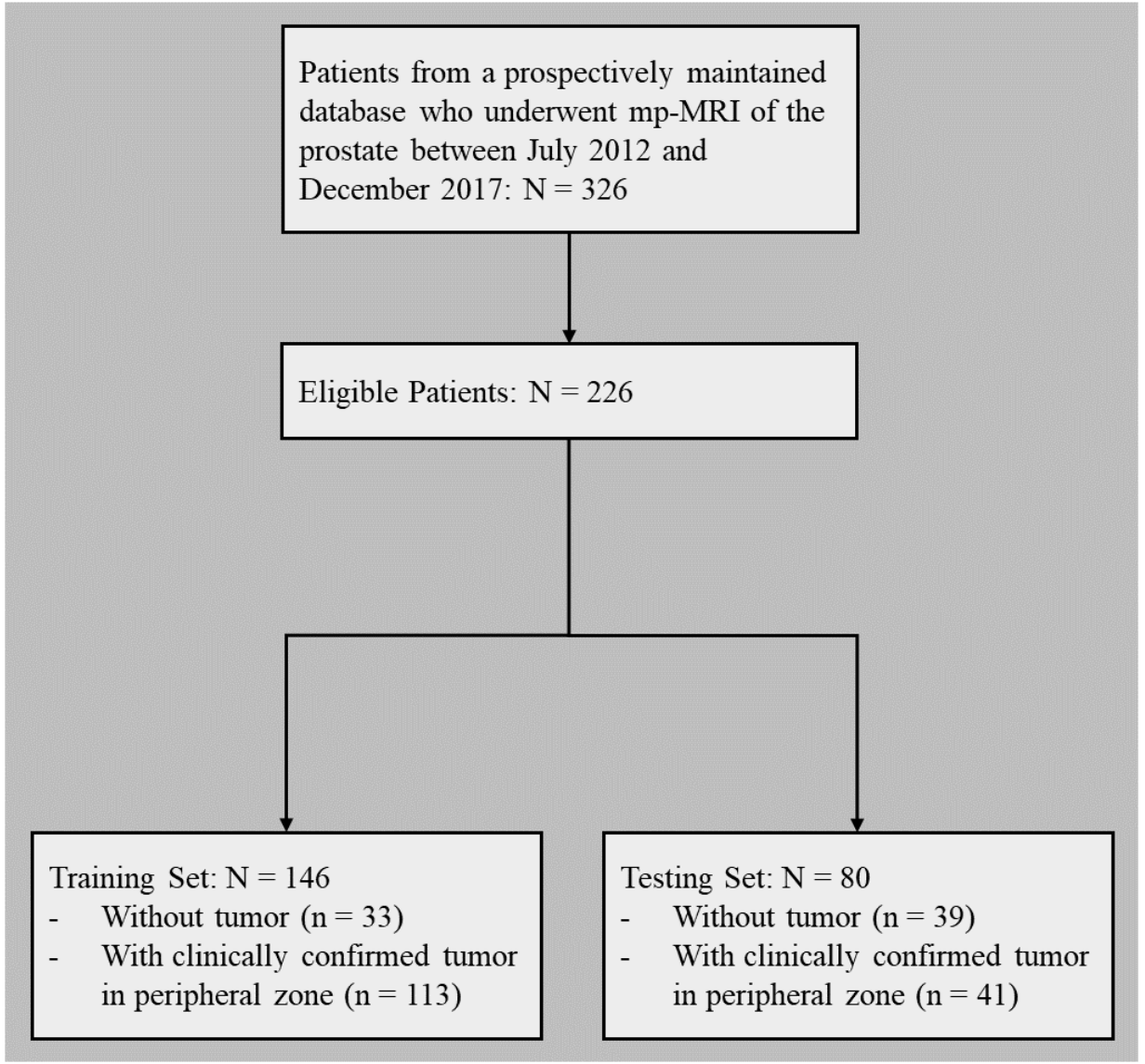


| MR Image | Segmented Region | DSC (%) |
|---|---|---|
| T2W | WG | 92.96 ± 7.77 |
| | CG | 91.07 ± 8.91 |
| | PZ | 86.22 ± 3.72 |
| ADC map | WG | 89.71 ± 8.89 |
| | CG | 86.33 ± 10.69 |
| | PZ | 83.30 ± 9.56 |

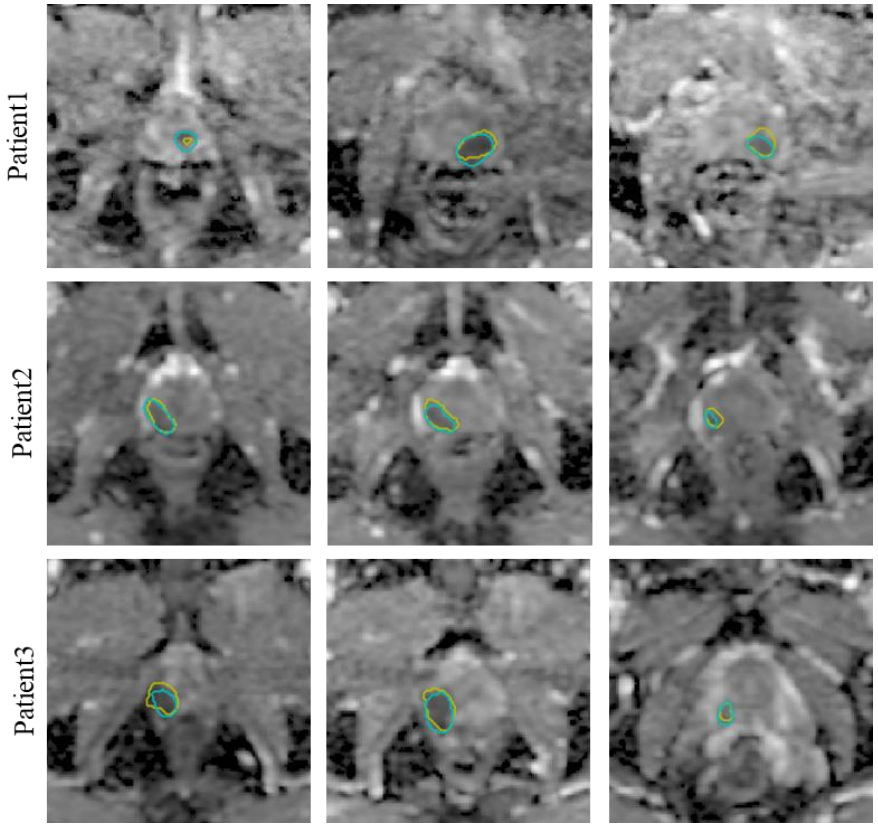
# The Block Diagram of Our Proposed Methodology for PCa Localization

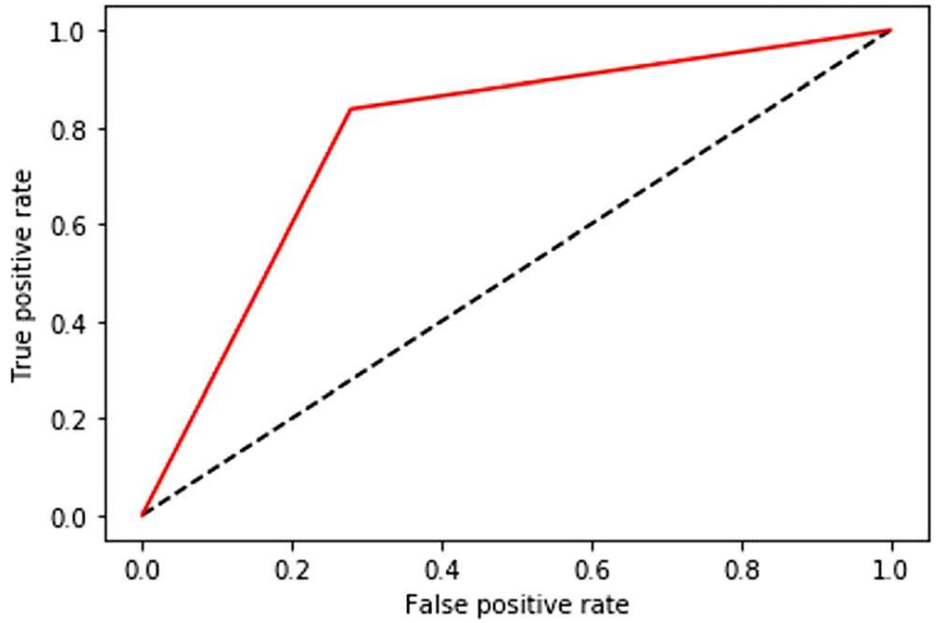


# Experimental Design



# PCa Localization Results




| DSC (%) | Sensitivity (%) | Specificity (%) | AUC |
|---|---|---|---|
| 86.72 ± 9.93 | 85.76 ± 23.33 | 76.44 ± 23.70 | 0.779 |

# Why Interpretability of Deep Learning Models Matters?

- When the AI is relatively weaker than the human and not yet reliably ‘deployable’, the goal of transparency and explanations is to identify the failure mode.

- ***When the AI is on par with humans and reliably ‘deployable’, the goal is to establish appropriate trust and confidence in users.***

- When the AI is significantly stronger than humans, the goal of the explanations is in machine teaching i.e., teaching humans how to take better decisions.

# Interpretability

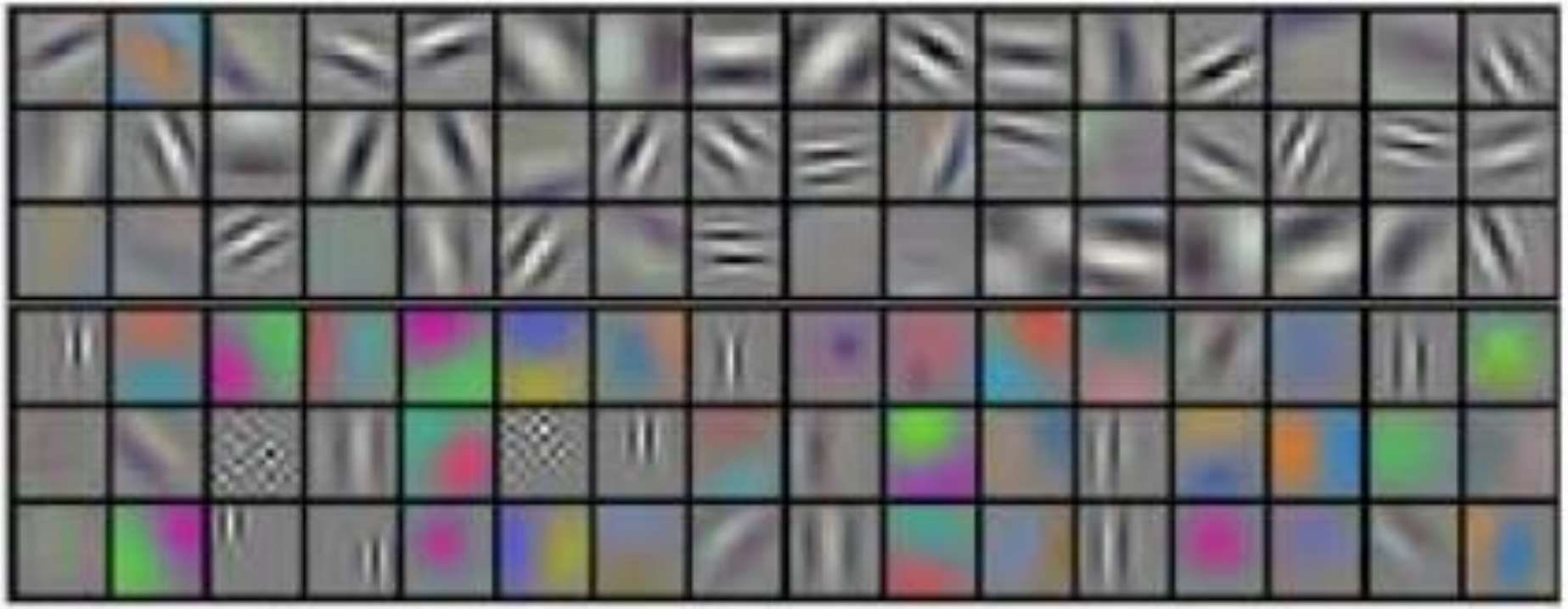
- When the AI algorithms are reliably deployable, it is crucial to establish appropriate trust and confidence in users.
- Interpretability of AI for medical applications, builds trust, and moves towards their successful integration in our daily lives.
- Interpretability = Feature selection:  the process of reducing the number of input variables when developing a predictive model.

**Criteria for feature selection?**

***"Good feature subsets contain features highly correlated with the class, yet uncorrelated with each other"*** Hall (1999).

# Interpretability in Image Analysis

Interpretability in image analysis (object detection, segmentation, and classification):



An example of feature map

# Grad-CAM [D. Mishra, 2019, https://towardsdatascience.com/]

***Where*** *the network is "looking" in the input image*

Observations:

- Convolutional feature map retain spatial information (which is lost in fully-connected layers)
- Each kernel represents some visual patterns
- Each pixel of the feature map indicates whether the corresponding kernel's visual pattern exists in its receptive fields.
- Last Convolutional Layer can be thought as the features of a classification model.

$y^c = f(A^1,...,A^k)$

# Grad-CAM Cont.

- Visualization of the final feature map ($A^k$) will show the discriminative region of the image.
- Simplest summary of all the $A^k$,k=1,...,K would be its linear combinations with some weights.
- Some feature maps would be more important to make a decision on one class than others, so weights should depend on the class of interest.

$$L^c_{Grad-CAM} \sim= \sum_{k=1}^{K} \alpha^c_k A^k \in \mathbb{R}^{u \, x \, v}$$

- So, the question is, what the weights should be?

# Weight Calculation

- The gradient of the cth class score with respect to feature maps $A^k$ measures linear effect of (i,j)th pixel point in the kth feature map on the cth class score.

$$\frac{dy^c}{dA^k_{i,j}}$$

- So averaging pooling of this gradient across i and j explains the effect of feature map k on the cth class score.

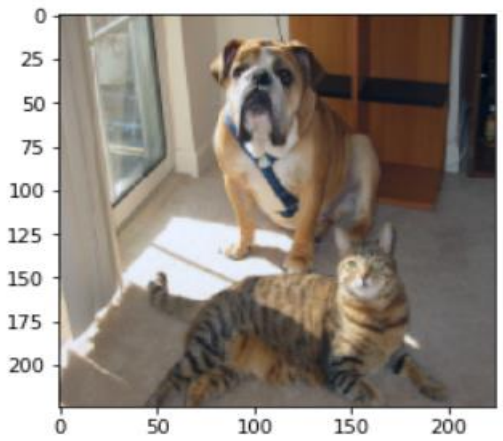
$$\alpha^c_k = \frac{1}{uv} \sum_{i=1}^{u} \sum_{j=1}^{v} \frac{dy^c}{dA^k_{i,j}}$$

- Grad-CAM propose to use this averaged gradient score as a weights for feature map.

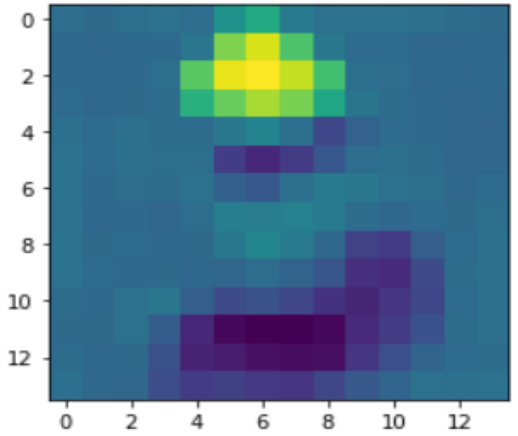
# Weight Calculation Cont.

- ReLU is applied to the linear combination of maps because we are only interested in the features that have a positive influence on the class of interest, i.e., pixels whose intensity should be increased in order to increase $y^c$.

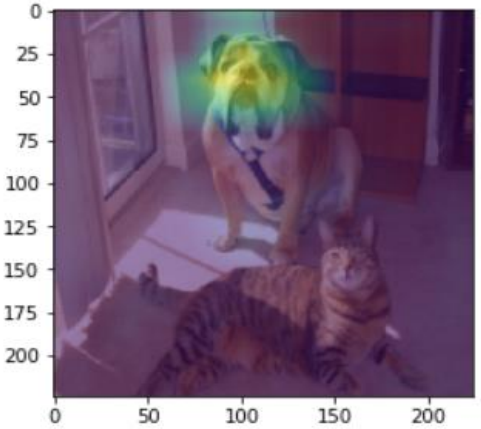
- Finally, we upsample the class activation map to the size of the input image to identify the image regions most relevant to the particular category.



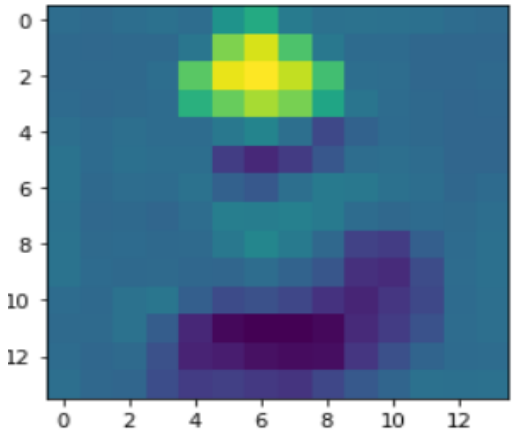
Input Image



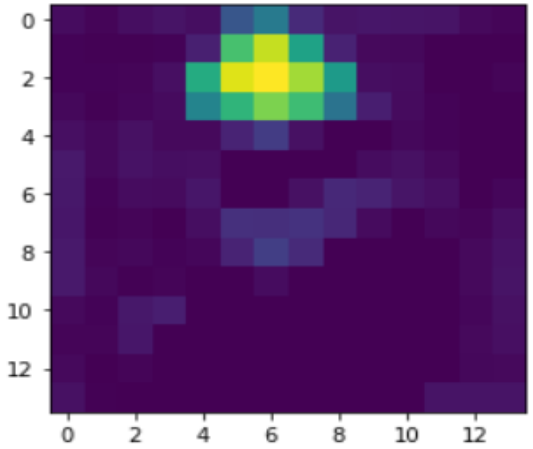
Class Discriminative Map after ReLU



Final Saliency Map with Bull Dog as Target Class



Class Discriminative Map



Class Discriminative Saliency Map after Normalization

[Retrieved from: towardsdatascience.com/demystifying-convolutional-neural-networks-using-gradcam-554a85dd4e48]

# Why variety of ML and DL-based algorithms?

No Free Lunch Theorems in artificial intelligence science:

*There is no perfect algorithm that works equally well for all tasks. Certain classes of algorithms have no “best” algorithm because on average, they’ll all perform about the same. Therefore, there is a need for a variety of tools.*

# Thank you!

# Any Question?