Machine Learning – CS74020
Spring 2025 – GC CUNY

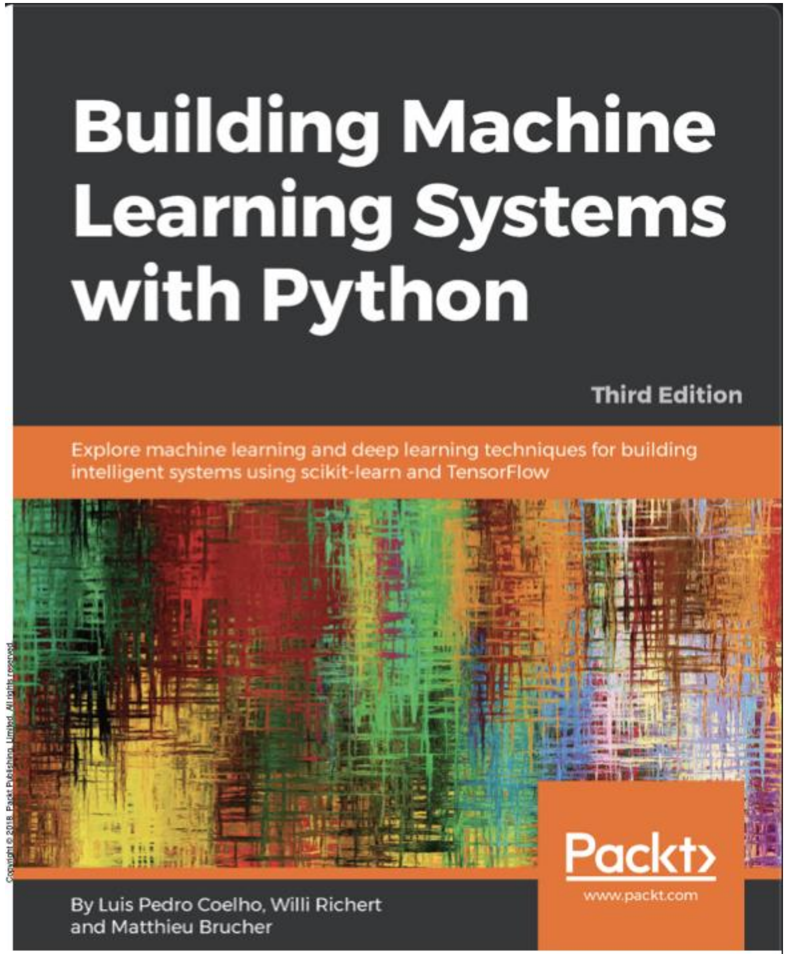
Week 01
Introduction to Machine Learning and Class Mechanics

Pegah Khosravi, PhD
Assistant Professor of Biomedical AI
New York City College of Technology (City Tech)
Faculty Member, Biology and Computer Science, CUNY Graduate Center

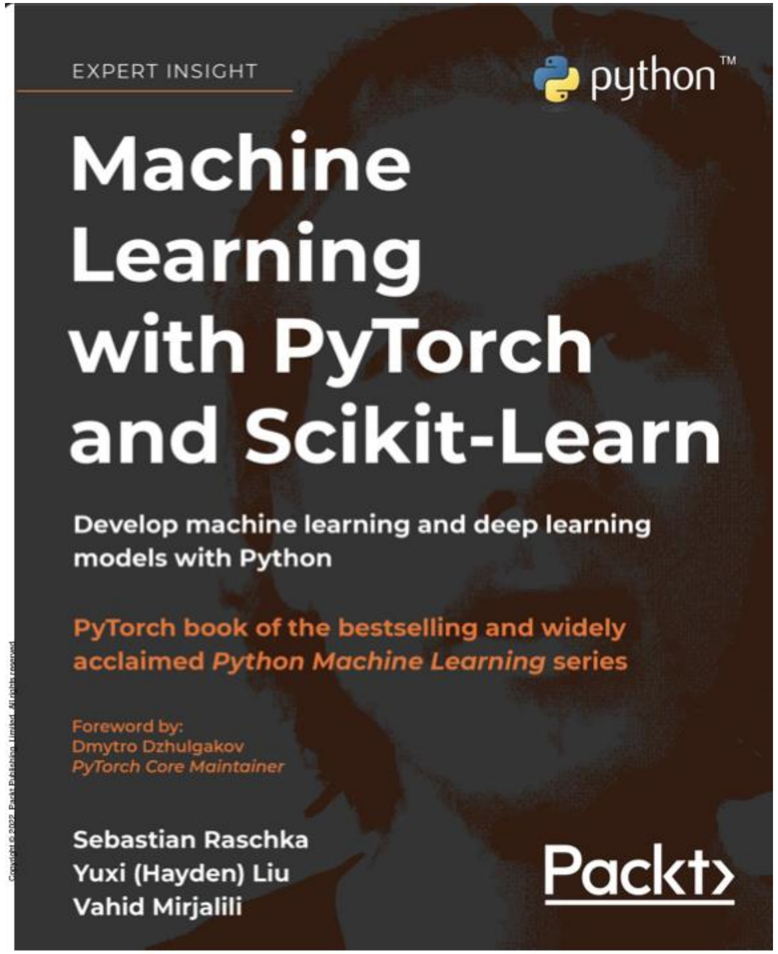
In this session, we will cover:

- Class Mechanics
- The three main types of Machine Learning (ML):
  - Supervised Learning
  - Unsupervised Learning
  - Semi-supervised Learning
- Generative vs. Discriminative models
- The basic workflow of ML projects
- Hands-on: Loading and exploring a dataset using Python

Textbooks



Pedro, Coelho, Luis, et al. Building Machine Learning Systems with Python : Explore Machine Learning and Deep Learning Techniques for Building Intelligent Systems Using Scikit-Learn and TensorFlow, 3rd Edition, Packt Publishing, Limited, 2018. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/citytech-ebooks/detail.action?docID=5485017
Created from citytech-ebooks on 2024-12-01 14:04:30.



Raschka, Sebastian, et al. Machine Learning with Pytorch and Scikit-Learn : Develop Machine Learning and Deep Learning Models with Python, Packt Publishing, Limited, 2022. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/citytech-ebooks/detail.action?docID=6870905.
Created from citytech-ebooks on 2024-12-01 13:57:51.

Grading Policies

ASSIGNMENT		DESCRIPTION	POINTS
Lab - Lecture	In class participation	Active participation in coding and analysis	15%
	Final Project	Comprehensive project due at the end of the course	15%
	Quizzes	Two quizzes assessing course topics	20%
	Exam 1	Midterm: Covers material from the first half of the course	25%
	Exam 2	Final: Covers material from the entire course	25%
Total			100%

Important Dates

- Class: In Person – GC6418 – Wednesdays 2 to 4 PM EST
- Quiz 1: Week 4 – 03/05
- Midterm: Week 7 – 03/19
- Quiz 2: Week 11 – 04/23
- Final Project: Week 14 – 05/14
- Final Exam: Week 15 – 05/21

What is Supervised Learning?

The model learns from labeled data (input-output pairs): Supervised learning is a core paradigm in ML, where the model is trained on a labeled dataset. Each training example consists of an input-output pair where:

- Input (X): A set of features (vectors) representing the data
- Output (Y): Corresponding target values (labels), which can be categorical or continuous

- Examples:
  - Predicting house prices (Regression)
  - Classifying emails as spam or not (Classification)

- Key Characteristics:
  - Training Data
  - Loss Function
    - Mean Squared Error (MSE) for regression
    - Cross-Entropy Loss for classification
  - Optimization: Techniques like Stochastic Gradient Descent (SGD)
  - Evaluation Metrics
    - Classification: Accuracy, Precision, Recall, F1-score, ROC-AUC
    - Regression: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE)

What is Unsupervised Learning?

The model identifies patterns in unlabeled data: The primary goal of unsupervised learning is to understand the underlying structure of the data, group similar data points, and reduce complexity where needed.

- Examples:
  - Clustering: K-Means, Hierarchical Clustering
  - Dimensionality Reduction: PCA, t-SNE

- Key Characteristics:
  - No Labels:
    - The dataset lacks predefined labels or target values.
    - The model explores the data's intrinsic patterns without guidance
  - Applications:
    - Clustering: Grouping data into similar clusters based on their features
    - Dimensionality Reduction: Simplifying datasets while retaining the most important information
  - Evaluation:
    - Metrics like silhouette score (for clustering) and explained variance (for dimensionality reduction) are commonly used.

What is Semi-supervised Learning?

Combines a small amount of labeled data with a large amount of unlabeled data: The central idea is that labeled data can guide the learning process, while the unlabeled data helps capture the broader structure of the dataset.

- Examples:
  - Label Propagation
  - Self-training
  - Co-training
  - Graph-based Method

- Key Characteristics:
  - Small Labeled Dataset
  - Large Unlabeled Dataset
  - Applications:
    - Medical imaging
    - Text classification
    - Speech recognition
  - Evaluation

Generative vs. Discriminative Models

- Generative Models and Discriminative Models are two types of machine learning models, but they focus on different tasks:

- Think of a detective solving a case:
  - Generative Model (storyteller): The detective tries to understand the full story of how the crime happened, recreating every detail (generating all possibilities).
  - Discriminative Model (decision-maker): The detective focuses only on the evidence at hand to decide whether the suspect is guilty or not (making a classification).

Aspect	Generative Models	Discriminative Models
Goal	Model P(X,Y)	Model P(Y|X)
Uses	Data generation, unsupervised learning	Classification, supervised learning
Examples	GANs, Naive Bayes, VAEs	Logistic Regression, SVMs, Neural Networks
Focus	How data is generated and structured	How to separate or classify data effectively

Generative Models

What they do?
Generative models aim to model the joint probability distribution P(X, Y) or the marginal probability P(X). This means they learn the structure of the data itself and how input features X and labels Y are related. By capturing this distribution, they can generate new plausible data samples.

Why they're useful?
They can generate new samples of data that resemble the original dataset. For example:
- A generative model trained on images of cats can create entirely new, realistic-looking images of cats.
- Generative models can also model uncertainty, learn variations in data, and fill in missing information.

Examples:
- **Naive Bayes**: A simple generative model for classification.
- **GANs (Generative Adversarial Networks)**: Used to create realistic images.
- **Variational Autoencoders (VAEs)**: Used for data generation and reconstruction.

Discriminative Models

What they do?
Discriminative models focus on modeling the conditional probability distribution $P(Y|X)$. Instead of learning the full data distribution, they focus on drawing decision boundaries between classes, making them better suited for classification and prediction tasks.

Why they're useful?
They are designed to efficiently separate or classify data and are often more accurate for tasks like supervised learning. Unlike generative models, they do not generate new data but focus on making precise classifications.

- For example, a discriminative model trained on cat and dog images will classify a new image as either a cat or a dog rather than generating a new cat image.

Examples:

- **Logistic Regression**: Classifies data into categories based on features.
- **Support Vector Machines (SVMs)**: Finds the best boundary to separate data.
- **Neural Networks (e.g., CNNs, RNNs)**: Excellent for tasks like image or text classification.

Machine Learning Workflow

A typical ML project follows these steps:

1. **Data Collection:** Gather data relevant to the problem.
2. **Preprocessing:** Clean and prepare the data.
3. **Model Training:** Use algorithms to find patterns in the data.
4. **Model Evaluation:** Assess performance on unseen data.
5. **Deployment:** Use the model in a real-world application.

A Deep Learning Approach to Diagnostic Classification of Prostate Cancer Using Pathology–Radiology Fusion

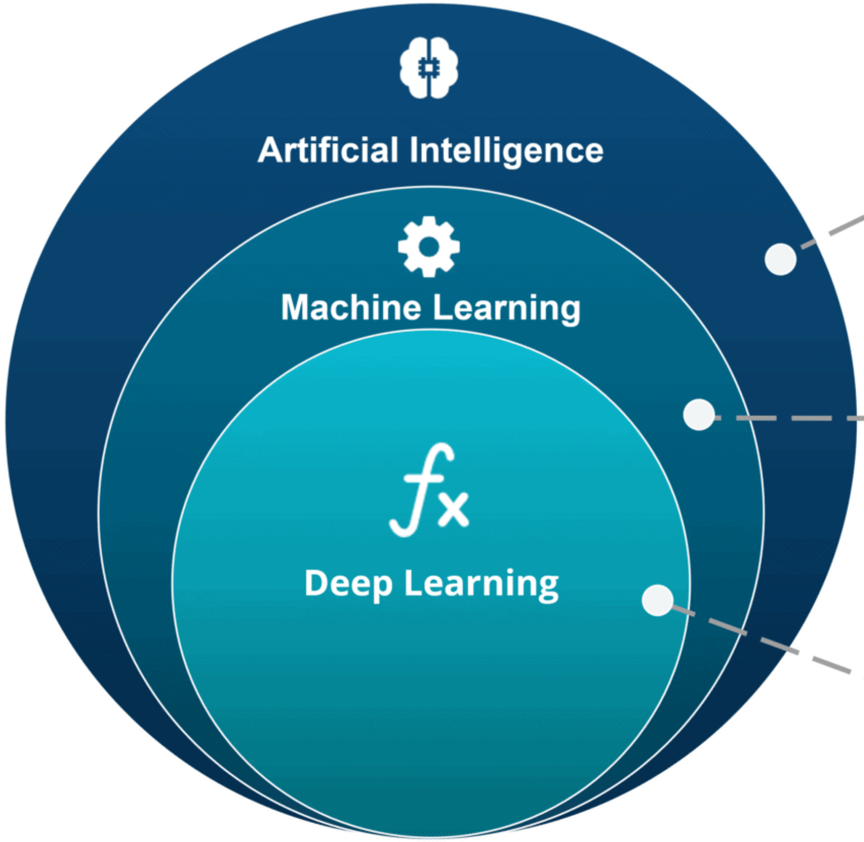
Pegah Khosravi, PhD
Assistant Professor
New York City College of Technology

02/05/2025

Outline

- AI, ML, DL
- CNN models
- Data and labels
- Training a CNN

AI - ML - DL



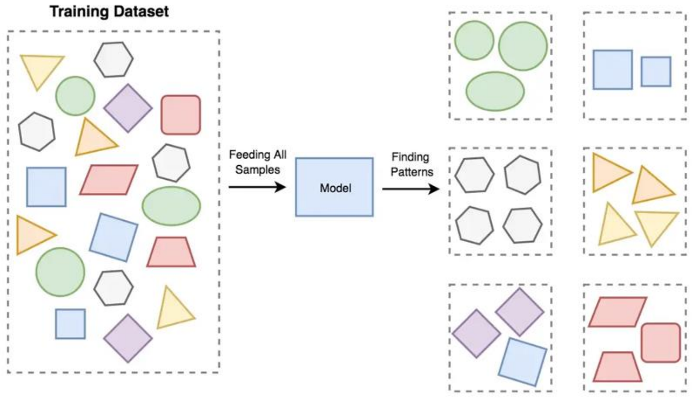
ARTIFICIAL INTELLIGENCE
A technique which enables machines to mimic human behaviour

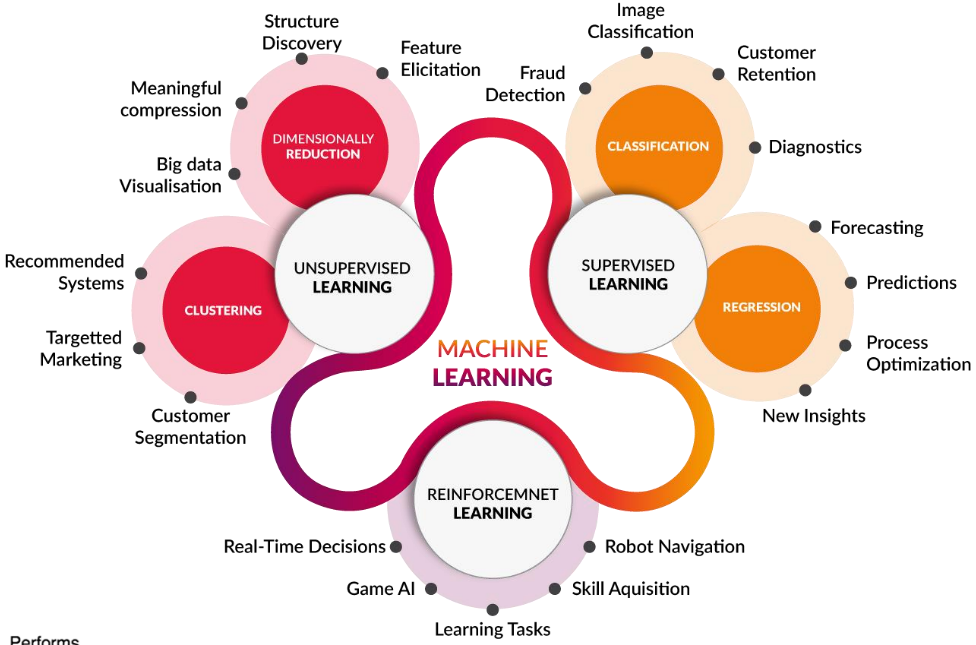
MACHINE LEARNING
Subset of AI technique which use statistical methods to enable machines to improve with experience

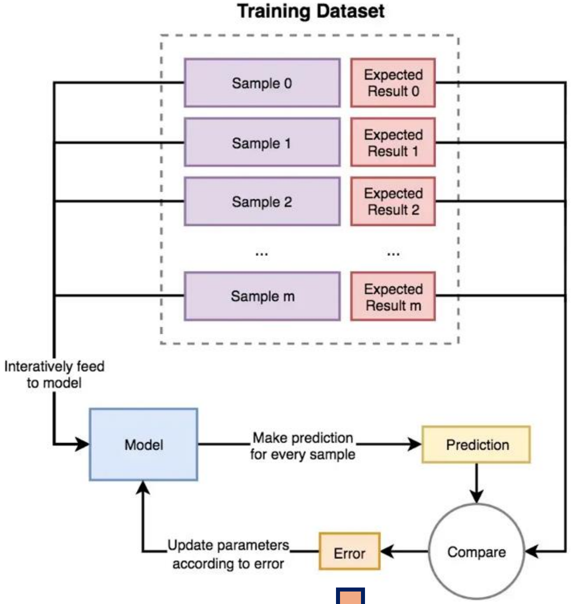
DEEP LEARNING
Subset of ML which make the computation of multi-layer neural network feasible

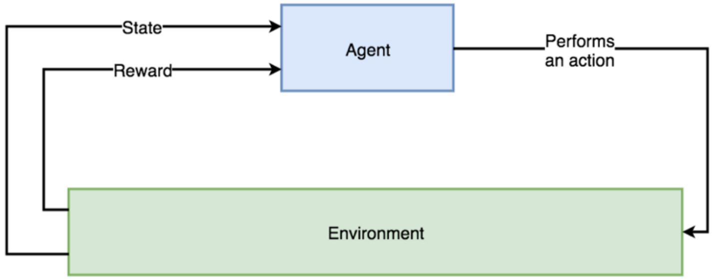
Atul, 2018, AI vs machine learning vs deep learning

Machine Learning Methods









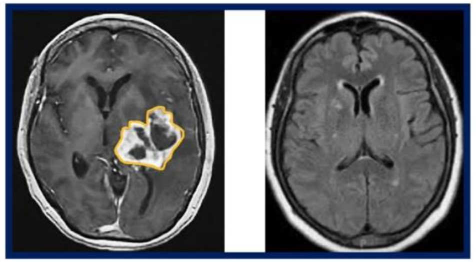
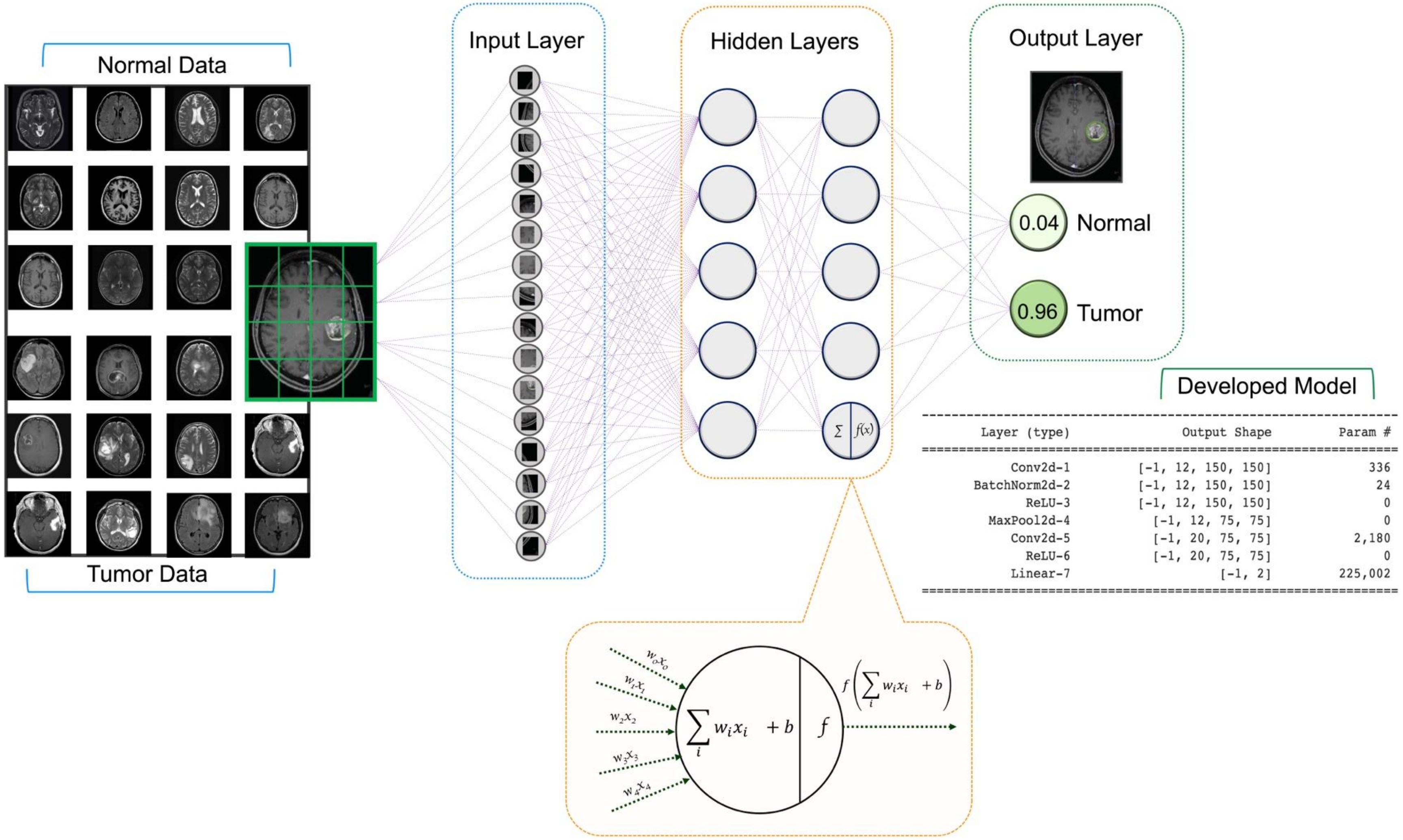


Image with Tumor

Image without Tumor

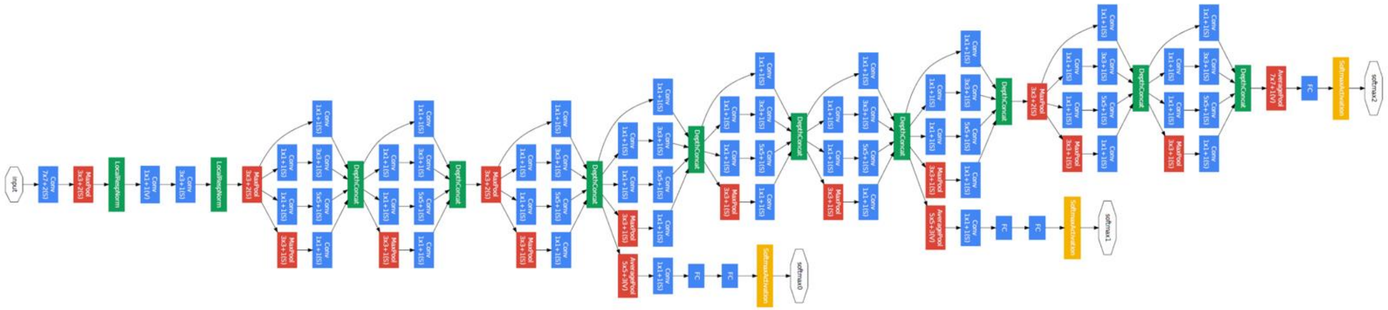
https://medium.com/towards-data-science/coding-deep-learning-for-beginners-types-of-machine-learning-b9e651e1ed9d

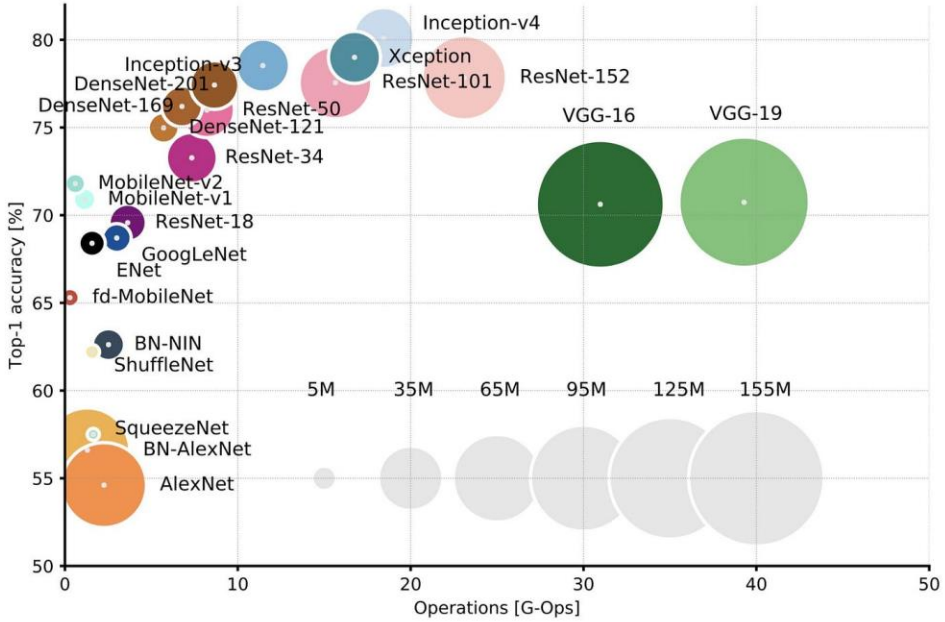
Deep Neural Network Algorithm



Khosravi, 2023, Artificial intelligence in neuroradiology: a scoping review of some ethical challenges

ILSVRC





ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

Annual competition between 2010 and 2017: The goal of the challenge was to both promote the development of better computer vision techniques and to benchmark the state of the art.

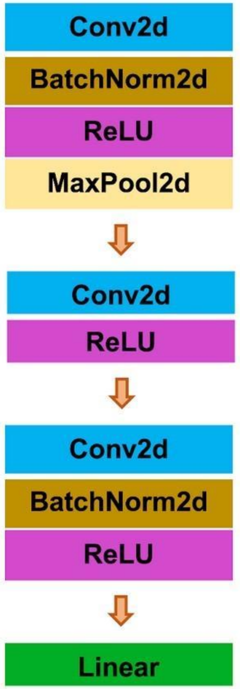
https://medium.com/@willwang1212/an-oversimplified-history-of-machine-learning-f5e171b2f6fa
https://medium.com/towards-data-science/neural-network-architectures-156e5bad51ba

Brain Tumor Detection – Training from the Scratch

```
[ ] from torchsummary import summary

    summary(model, (3,150,150))

----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv2d-1         [-1, 12, 150, 150]             336
       BatchNorm2d-2         [-1, 12, 150, 150]              24
              ReLU-3         [-1, 12, 150, 150]               0
         MaxPool2d-4           [-1, 12, 75, 75]               0
            Conv2d-5           [-1, 20, 75, 75]           2,180
              ReLU-6           [-1, 20, 75, 75]               0
            Conv2d-7           [-1, 32, 75, 75]           5,792
       BatchNorm2d-8           [-1, 32, 75, 75]              64
              ReLU-9           [-1, 32, 75, 75]               0
           Linear-10                    [-1, 2]         360,002
================================================================
Total params: 368,398
Trainable params: 368,398
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 0.26
Forward/backward pass size (MB): 12.53
Params size (MB): 1.41
Estimated Total Size (MB): 14.19
----------------------------------------------------------------
```



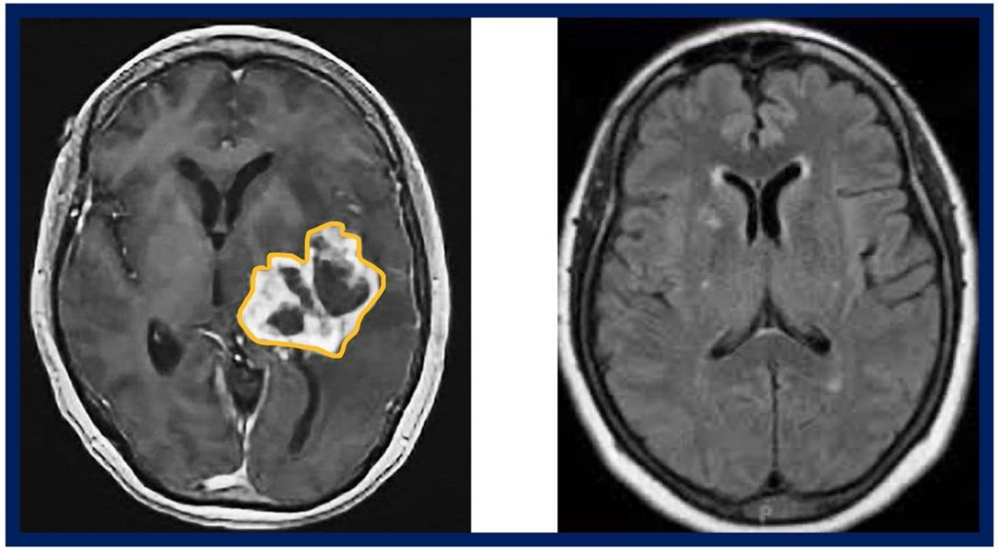


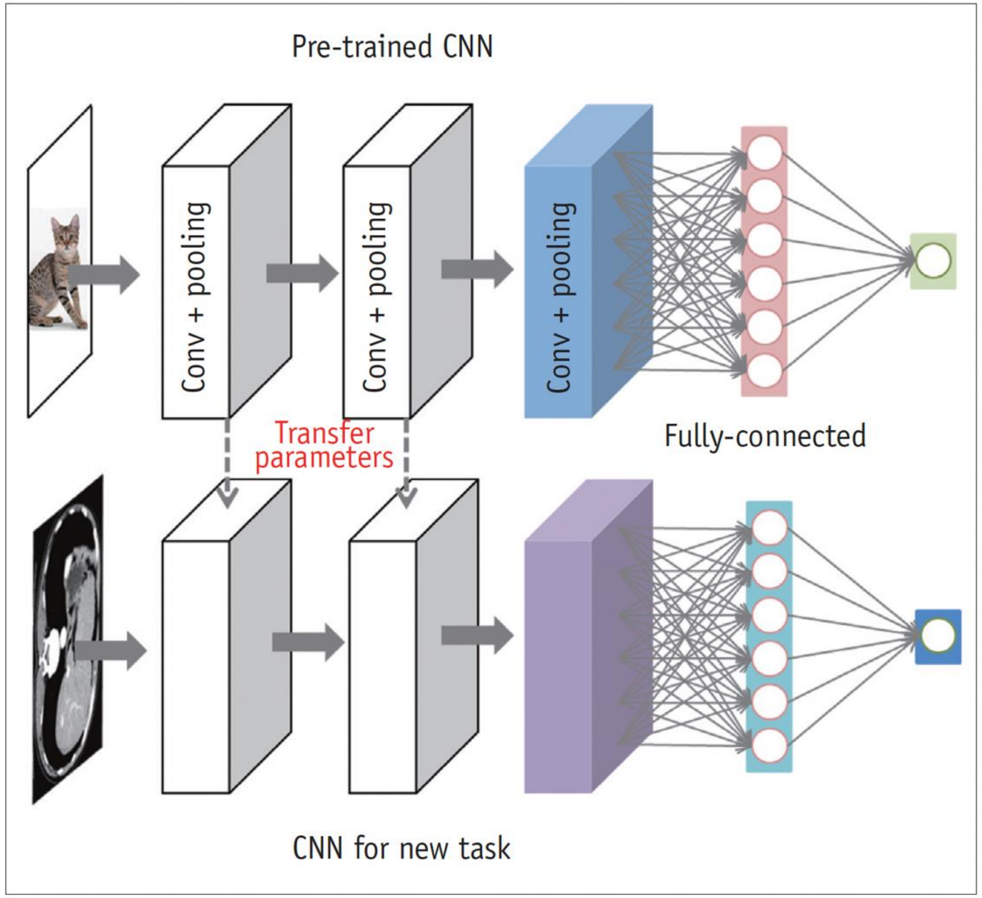
Image with Tumor

Image without Tumor

Performance: AUC = 0.83
7200 images from public data

https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection
https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

Transfer Learning

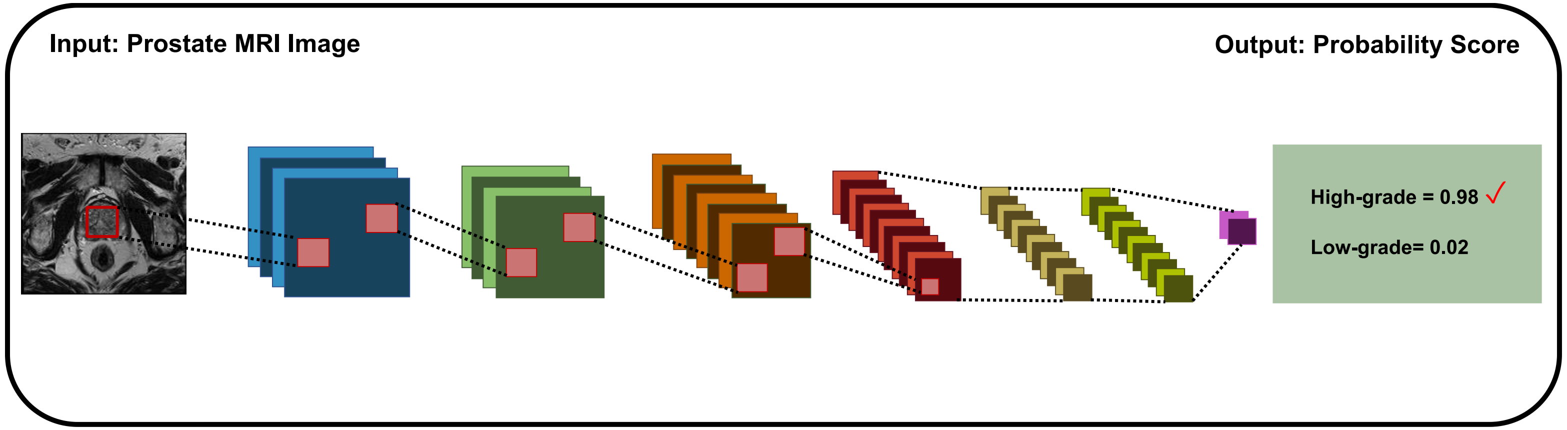


Transfer learning refers to the technique of leveraging **pre-trained deep learning models** on a large dataset and applying them to a new task or domain with limited labeled data.
TL addresses the limitation of large annotated datasets in medical image analysis by using pre-trained models, which have already learned rich feature representations from a different dataset.
These pre-trained models have captured general patterns and structures that are useful for various visual recognition tasks.
By using transfer learning, the **lower-level features** learned by the pre-trained model can be reused, while the **higher-level layers** can be fine-tuned or retrained on the new target dataset. T
his approach allows the model to learn specific features relevant to the new task with limited labeled data, improving its performance and generalization.

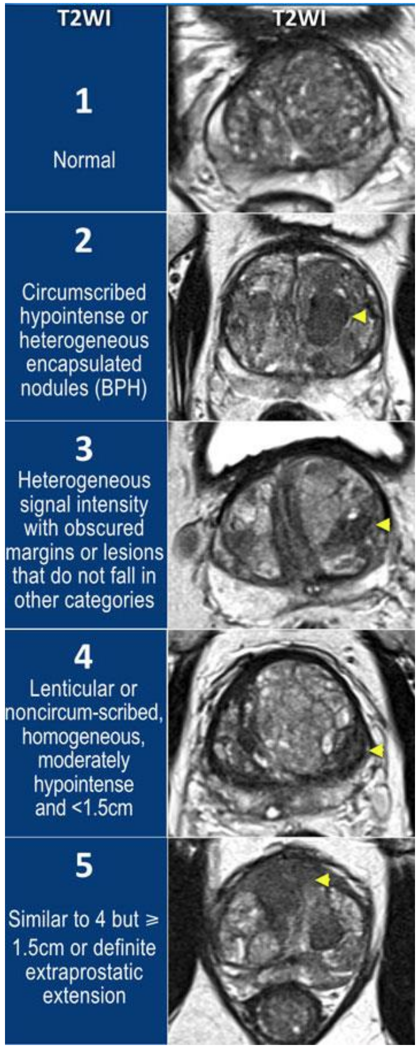
https://www.sciencedirect.com/science/article/abs/pii/B978032390534300010X

CNN

- A **convolutional neural network (CNN)** is a type of DL model that is primarily used for analyzing visual data such as images or videos. It is a specialized type of neural network designed to automatically and efficiently learn hierarchical patterns and features from input data.
- **CNN** is its ability to perform convolution operations, which involve sliding small filters (also known as **kernels**) over the input data to extract local features.
- We hypothesize that **CNN** can be used to predict prostate cancer aggressiveness using MR images only that are labeled based on histopathology information.
- Distinguishing patients with **high-risk** (tumor tissue growing faster) and **low-risk** (tumor tissue growing slowly) forms of prostate cancer is important:
  - early detection of high-grade prostate cancer improves **survival rate**
  - accurate diagnosis prevents **over-treatment**



PI-RADS and Biopsy



TRADITIONAL GLEASON SCORE	NEW GRADING SYSTEM GROUP 1
GLEASON 3+3=6 Only individual discrete well-formed glands	GRADE 1
GLEASON 3+4=7 Predominantly well-formed glands with a lesser component of poorly-formed/fused/cribiform glands.	GRADE 2
GLEASON 4+3=7 Predominantly poorly-formed/ fused/cribriform glands with a lesser component of well-formed glands.	GRADE 3
GLEASON 4+4=8 Only poorly-formed/fused/cribriform glands or -Predominantly well-formed glands with a lesser component lacking or -Predominantly lacking glands with a lesser component of well-formed glands.	GRADE 4
GLEASON 9-10 Lacks gland formation (or with necrosis) with or without poorly-formed/fused/cribriform gland.	GRADE 5

- We hypothesize that trained CNN model can increases the accuracy of **PI-RADS** scoring for prostate cancer.
- The trained model integrates complementary information from **biopsy** reports and improves diagnosis beyond what is possible with MR images alone.

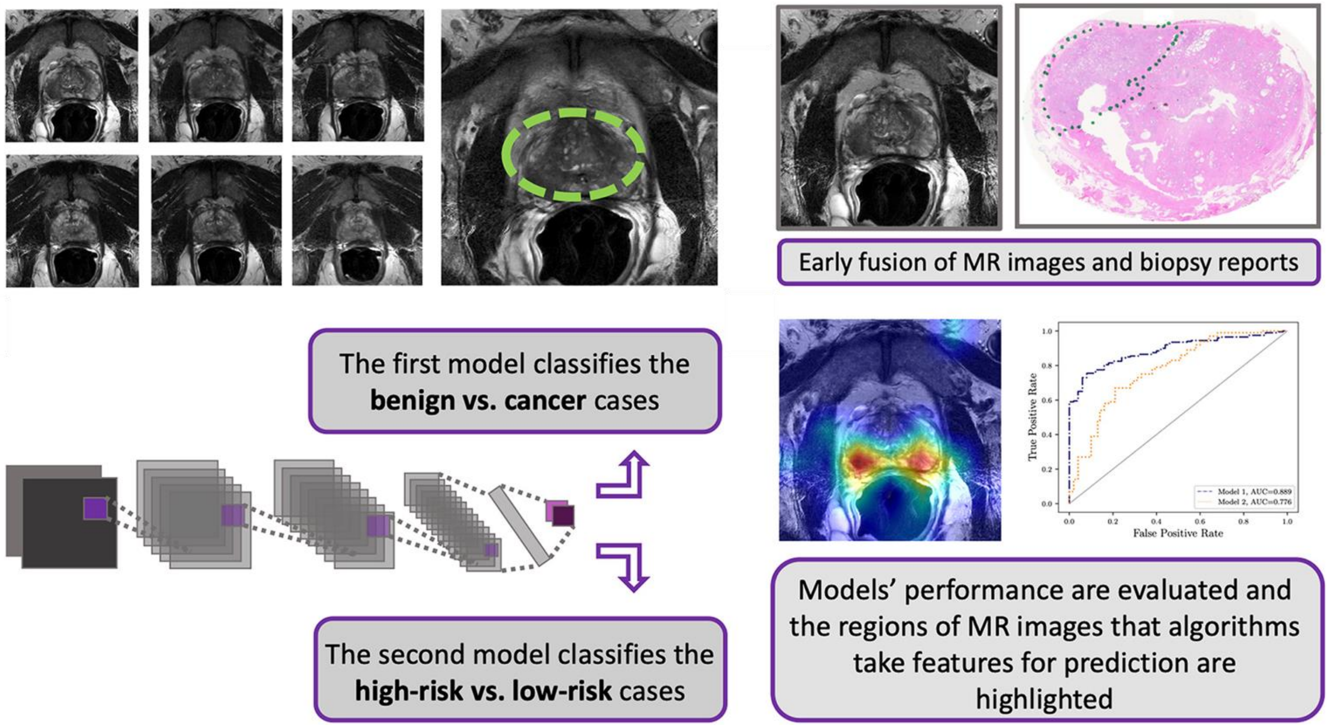
https://staging.radiologyassistant.nl/abdomen/prostate/prostate-cancer-pi-rads-v2
https://hudsonvalleypress.com/2019/01/02/cancer-scoring-method-may-underestimate-mortality-risk/

Multimodal Data Fusion

The aim of this study is to combine **Magnetic Resonance Imaging (MRI)** data with **pathology assessment** from **400 patients** with suspected **prostate cancer** to develop an **Artificial Intelligence** model **(AI-biopsy)** for the early diagnosis of prostate cancer.

The MR images were selected by expert radiologists from five different institutes

Two deep learning models were trained using MR images

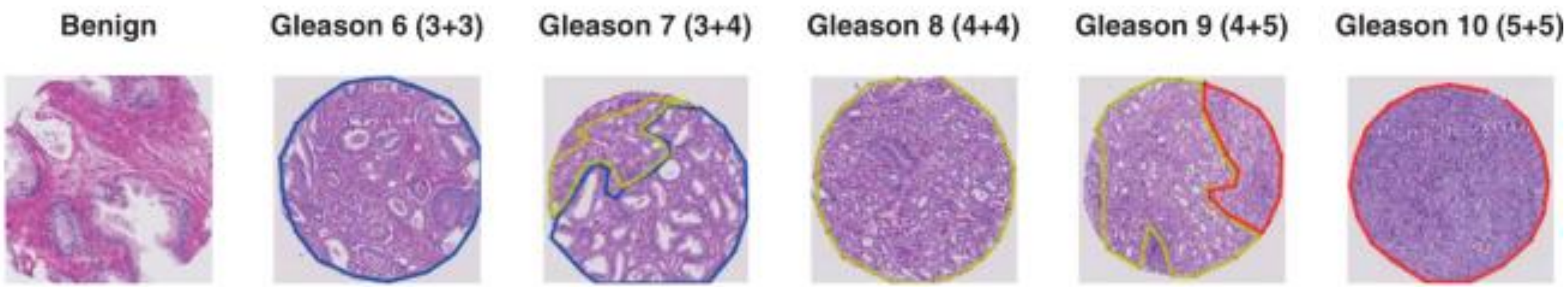


The MR images were labeled by biopsy reports instead of PI-RADS scores

A platform was developed to automatically distinguish cancer patients from benign patients and high-risk tumors from low-risk tumors

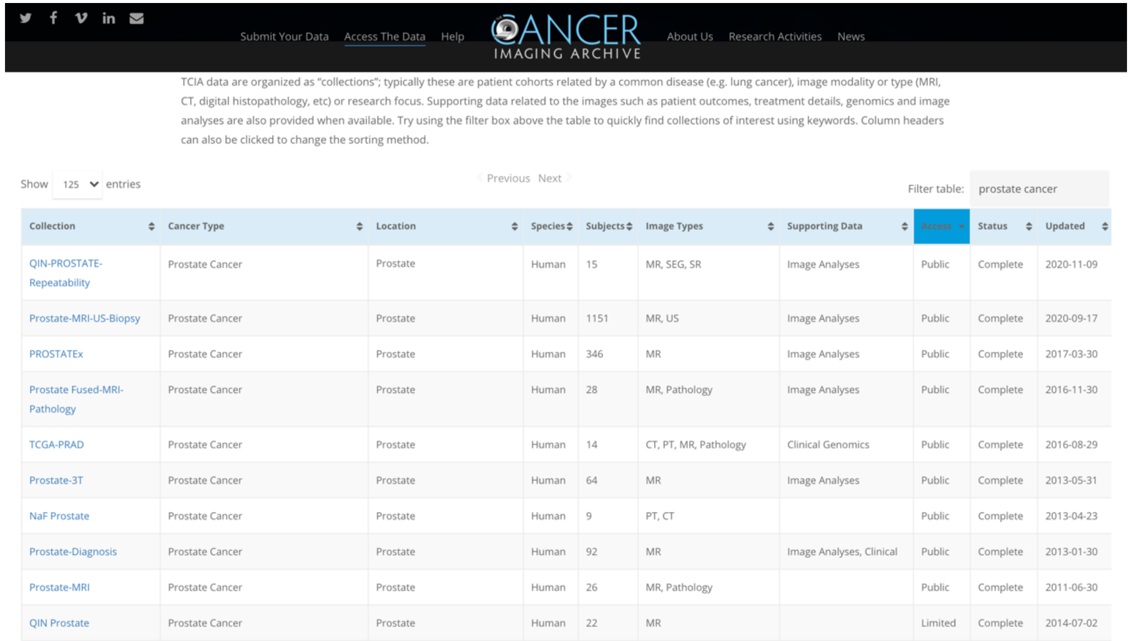
Grading system

Gleason Grade Group	Gleason Score	Combined Gleason Score	Risk
grade group1	3+3	6	low risk
grade group2	3+4	7	Intermediate risk closer to low risk
grade group3	4+3	7	Intermediate risk closer to high risk
grade group4	4+4,3+5,5+3	8	High risk
grade group5	4+5,5+4,5+5	9 and 10	High risk



National comprehensive cancer network (NCCN) guideline for prostate cancer

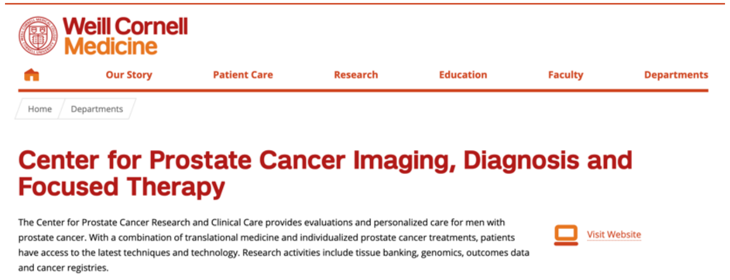
Public and in-house Data


Submit Your Data | Access The Data | Help | CANCER IMAGING ARCHIVE | About Us | Research Activities | News

TCIA data are organized as "collections"; typically these are patient cohorts related by a common disease (e.g. lung cancer), image modality or type (MRI, CT, digital histopathology, etc) or research focus. Supporting data related to the images such as patient outcomes, treatment details, genomics and image analyses are also provided when available. Try using the filter box above the table to quickly find collections of interest using keywords. Column headers can also be clicked to change the sorting method.

Show 125 entries | Previous Next | Filter table: prostate cancer

Collection	Cancer Type	Location	Species	Subjects	Image Types	Supporting Data	Access	Status	Updated
QIN-PROSTATE-Repeatability	Prostate Cancer	Prostate	Human	15	MR, SEG, SR	Image Analyses	Public	Complete	2020-11-09
Prostate-MRI-US-Biopsy	Prostate Cancer	Prostate	Human	1151	MR, US	Image Analyses	Public	Complete	2020-09-17
PROSTATEx	Prostate Cancer	Prostate	Human	346	MR	Image Analyses	Public	Complete	2017-03-30
Prostate Fused-MRI-Pathology	Prostate Cancer	Prostate	Human	28	MR, Pathology	Image Analyses	Public	Complete	2016-11-30
TCGA-PRAD	Prostate Cancer	Prostate	Human	14	CT, PT, MR, Pathology	Clinical Genomics	Public	Complete	2016-08-29
Prostate-3T	Prostate Cancer	Prostate	Human	64	MR	Image Analyses	Public	Complete	2013-05-31
NaF Prostate	Prostate Cancer	Prostate	Human	9	PT, CT		Public	Complete	2013-04-23
Prostate-Diagnosis	Prostate Cancer	Prostate	Human	92	MR	Image Analyses, Clinical	Public	Complete	2013-01-30
Prostate-MRI	Prostate Cancer	Prostate	Human	26	MR, Pathology		Public	Complete	2011-06-30
QIN Prostate	Prostate Cancer	Prostate	Human	22	MR		Limited	Complete	2014-07-02


Weill Cornell Medicine

Our Story | Patient Care | Research | Education | Faculty | Departments

Home / Departments

Center for Prostate Cancer Imaging, Diagnosis and Focused Therapy

The Center for Prostate Cancer Research and Clinical Care provides evaluations and personalized care for men with prostate cancer. With a combination of translational medicine and individualized prostate cancer treatments, patients have access to the latest techniques and technology. Research activities include tissue banking, genomics, outcomes data and cancer registries.

Visit Website

The Cancer Imaging Archive (TCIA)
https://www.cancerimagingarchive.net/collections/
PMID: 23884657

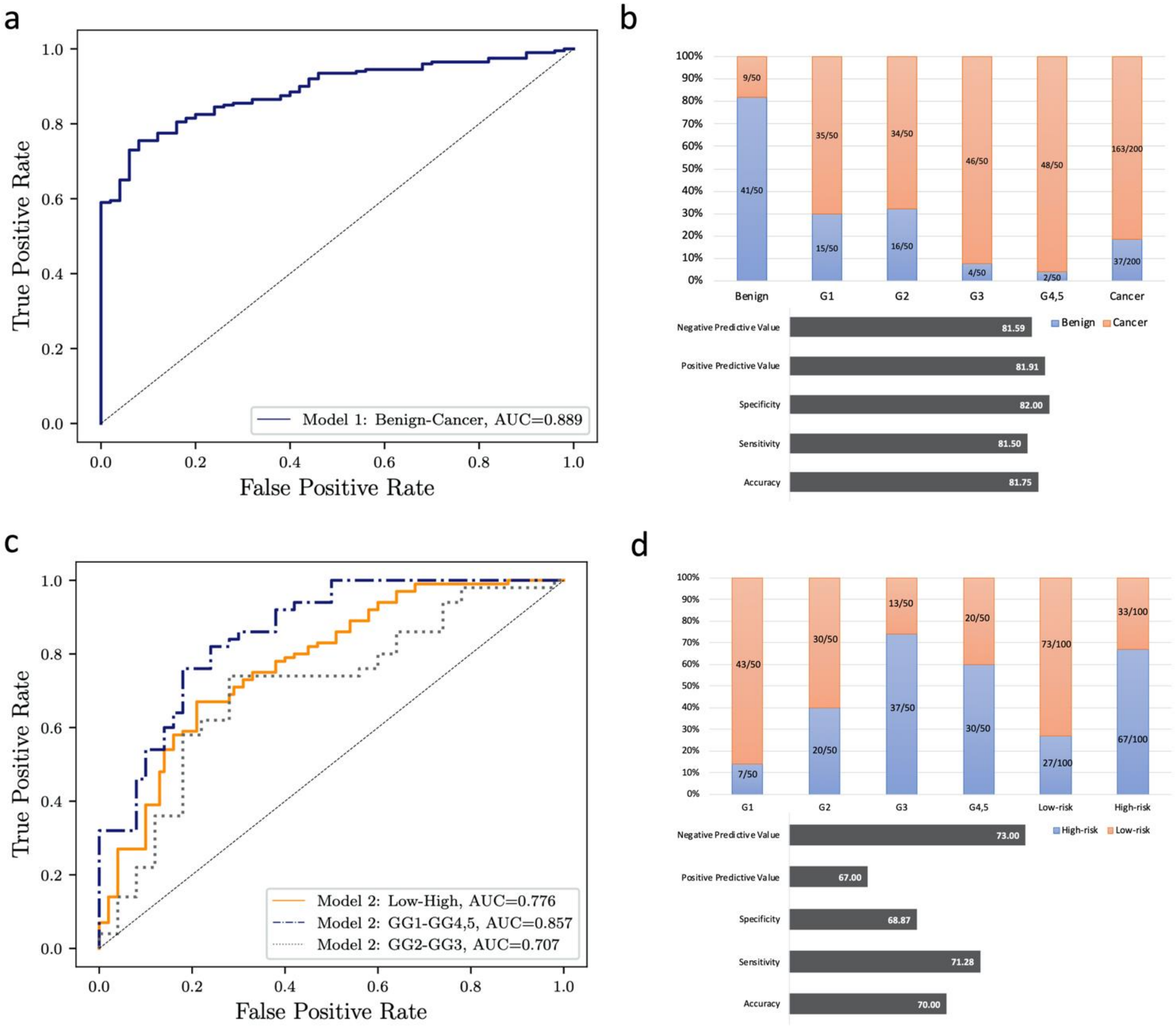
Data

database	Annotated patients	Annotation method	Cancer patients				Benign cases
			High-risk (GS ≥ 8) (GG=4 & GG=5)	Low-risk (GS = 6) (GG=1)	Intermediate-risk (GS = 7) (GG=2)	intermediate risk (GS = 7) (GG=3)	Benign
Cornell Medicine	228	Gleason Score	11	48	37	15	117
PROSTATEx	99	Grade group	13	29	38	19	0
PROSTATE-DIAGNOSIS	38	Pathology report	9	5	15	9	0
PROSTATE-MRI	26	Pathology slides	11	0	13	2	0
TCGA-PRAD	9	Gleason score	4	0	3	2	0
Total	400	Grade group Gleason score	48	82	106	47	117

Models

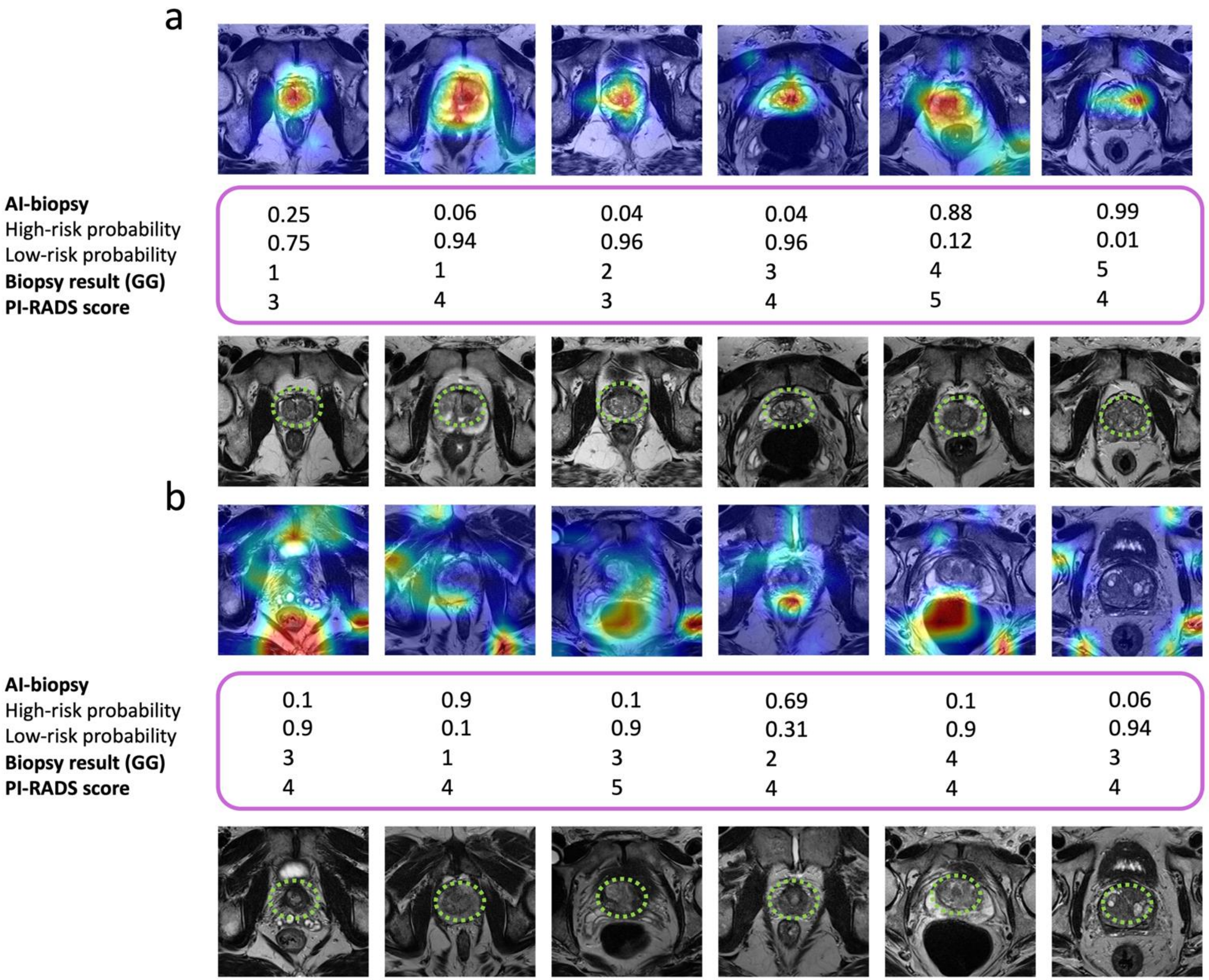
Model	Data resources	Number of patients with cancerous tumorin training and validation sets	Number of patients with benign tumor in training and validation sets	Number of patients in test set	Total number of patients in test set
Model1: Benign vs. Cancer	In-house and public	75 patients (37 GG=3, 38 GG=4 and GG=5)	107 patients (benign)	10 Benign 10 GG = 1 10 GG = 2 10 GG = 3 10 GG = 4&5	Five times cross validation of 50 patients
Model	Data resources	Number of patients with high-risk tumor in training and validation sets	Number of patients with low-risk tumor in training and validation sets	Number of patients in test set	Total number of patients in test set
Model2: High-risk vs. Low-risk	In-house and public	75 patients (37 GG=3, 38 GG=4 and GG=5)	168 patients (72 GG=1 and 96 GG=2)	10 GG = 1 10 GG = 2 10 GG = 3 10 GG = 4&5	Five times cross validation of 40 patients

Outline



Cohen's kappa		
PIRAD	0.19	0.1-0.2 slight
Algorithm	0.47	0.4-0.6 moderate

CAM



Zhou, et al., A. Learning Deep Features for Discriminative Localization. *2016*

Paper

JMRI Journal of Magnetic Resonance Imaging
ISMRM
ONE COMMUNITY FOR CLINICIANS AND SCIENTISTS

Original Research | Open Access

A Deep Learning Approach to Diagnostic Classification of Prostate Cancer Using Pathology–Radiology Fusion

Pegah Khosravi PhD, Maria Lysandrou BS, Mahmoud Eljalby MMS, Qianzi Li BA, Ehsan Kazemi PhD, Pantelis Zisimopoulos MS, Alexandros Sigaras MS, Matthew Brendel MEng, Josue Barnes MS, Camir Ricketts PhD, Dmitry Meleshko MS, Andy Yat RT, Timothy D. McClure MD, Brian D. Robinson MD, Andrea Sboner PhD, Olivier Elemento PhD, Bilal Chughtai MD, Iman Hajirasouliha PhD
... See fewer authors

First published: 14 March 2021 | https://doi.org/10.1002/jmri.27599 | Citations: 1

Thanks
Questions?