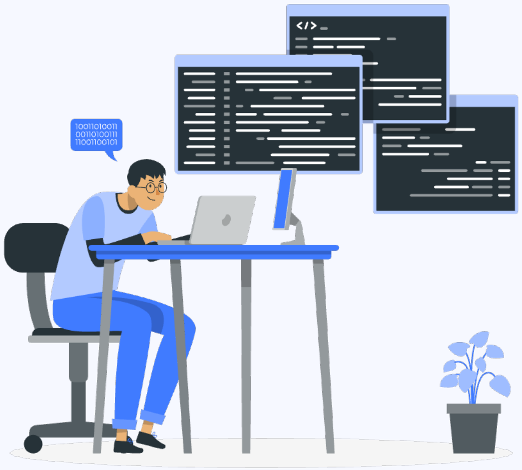
From Tiny-AI To Lite-AI in Edge Computing

Progress toward the middle

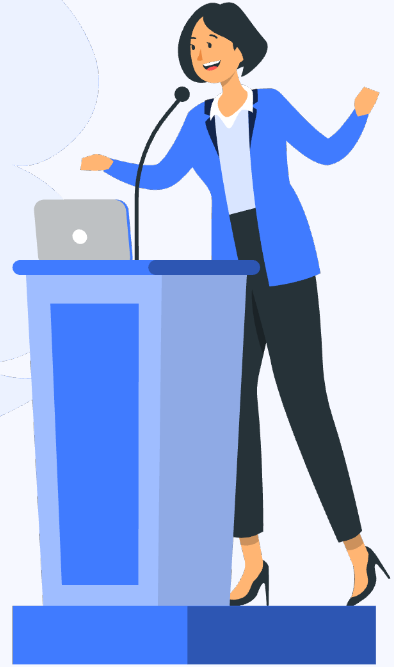


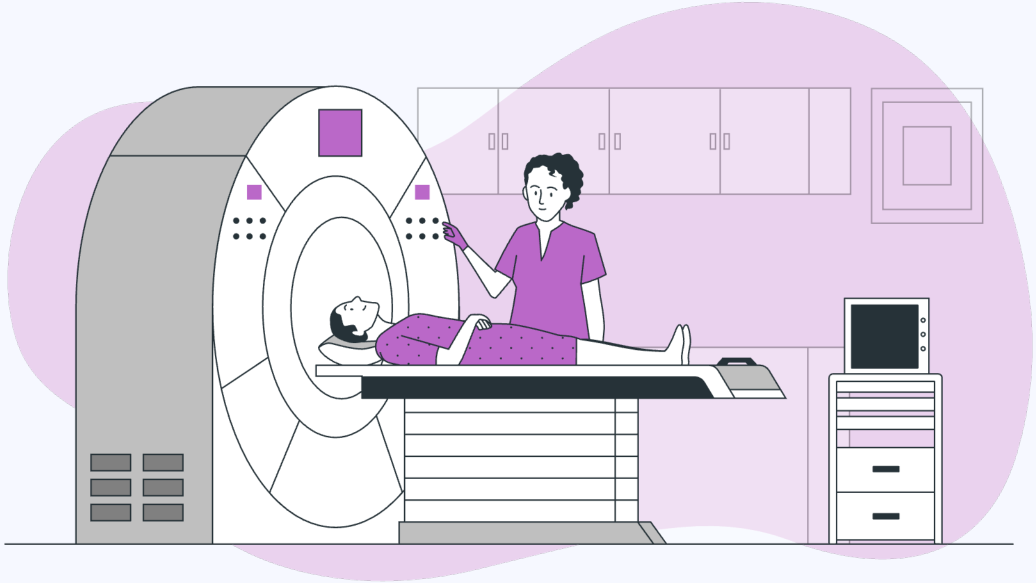
Biography

- Dr. Adli Md Ali is a:
  - Assistant Professor @ IIUM, Kulliyyah of Science, Dept. Physics
  - Research Fellow @ Microwave Research Institute, UiTM
- Research:
  - Development of resource-efficient ML models
  - Ethics in clinical Ai
  - ML development for sensor - just recently

- Skill:
  - A coder → Python , ML and ANN
  - High performance computing
  - Distributed computing system
- Dislike:
  - Writing paper , grant report writing

TABLE OF CONTENTS





01

WHY Do we need Lite-Ai

The Coming Clinical - Ai Revolution

MARKETS BUSINESS INVESTING TECH POLITICS CNBC TV INVESTING CLUB PRO

TECHNOLOGY EXECUTIVE COUNCIL

The A.I. revolution in health care is coming

PUBLISHED WED, JUL 12 2023·1:07 PM EDT

Rachel Curry
@WRITINGSOFRACH

SHARE

THE TIMES
MONDAY AUGUST 14 2023

Log in Subscribe Search

RACHEL SYLVESTER

The AI revolution can put patients at the centre of NHS

Self-triage, augmented reality surgery and robot helpers are some of the Israeli innovations we would do well to replicate

Rachel Sylvester | Thursday May 25 2023, 5.00pm, The Times

Subscribe to newsletters

Forbes

FORBES > SMALL BUSINESS

The Coming AI Revolution In Home Care

Josh Klein Forbes Councils Member
Forbes Business Council COUNCIL POST | Membership (Fee-Based)

Unparallel Accuracy?

SEARCH FORTUNE Subscribe Now SIGN IN

TECH · ALZHEIMER'S

Researchers used artificial intelligence to detect Alzheimer's risk with over 90% accuracy and could transform how medicine is practiced

Google's AI for medicine shows clinical answers more than 90 pct accurate

Bloomberg

Published: 13 July ,2023: 12:23 AM GST
Updated: 13 July ,2023: 12:26 AM GST

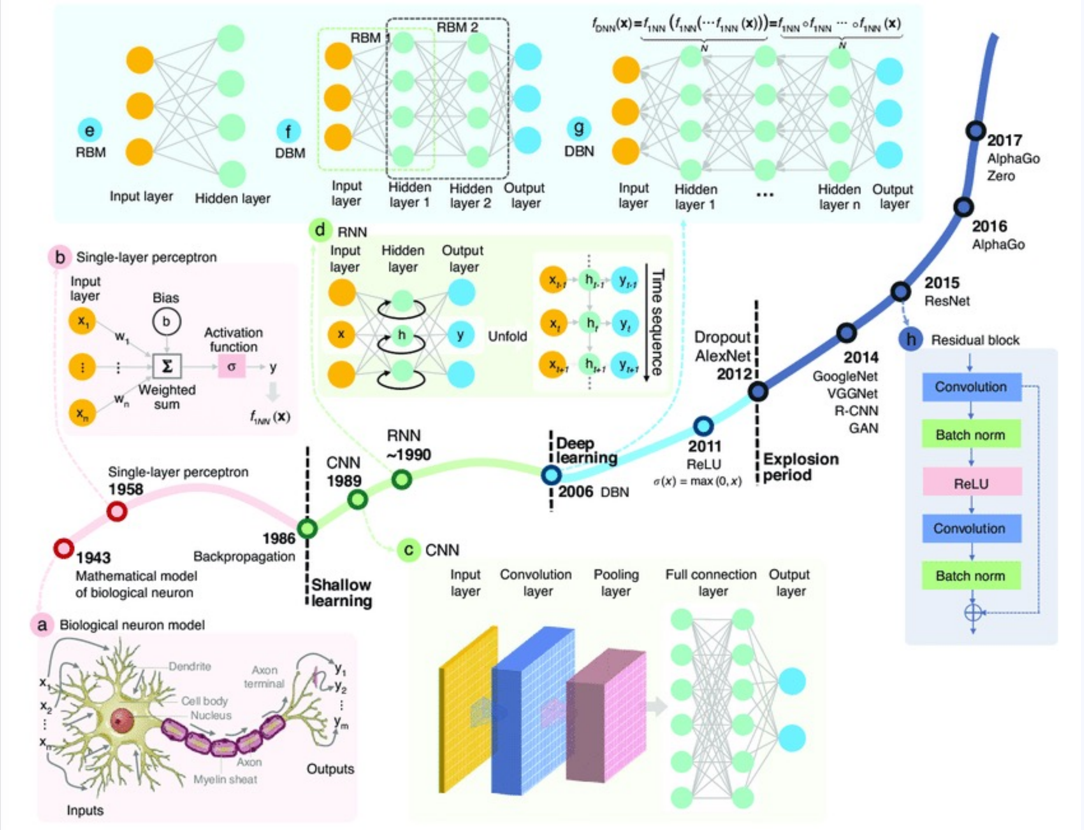
AI In Healthcare: New AI Model Diagnosed Heart Attacks With 99% Accuracy; May Help Doctors In Future

Currently, the tool is undergoing clinical trials in Scotland.

By Vikas Yadav Sat, 13 May 2023 12:27 AM (IST) Source:JND

Progress of **ANN**

From a shallow – simple CNN to multilayer deep complex model



Zuo, Chao & Qian, Jiaming & Feng, Shijie & Yin, Wei & Li, Yixuan & Fan, Pengfei & Han, Jing & Qian, Kemao & Chen, Qian. (2022). Deep learning in optical metrology: a review. Light: Science & Applications. 11. 39. 10.1038/s41377-022-00714-x.

Progress of ANN

Number of parameter per-model increase significantly in just few years

7 x

↑ >10%

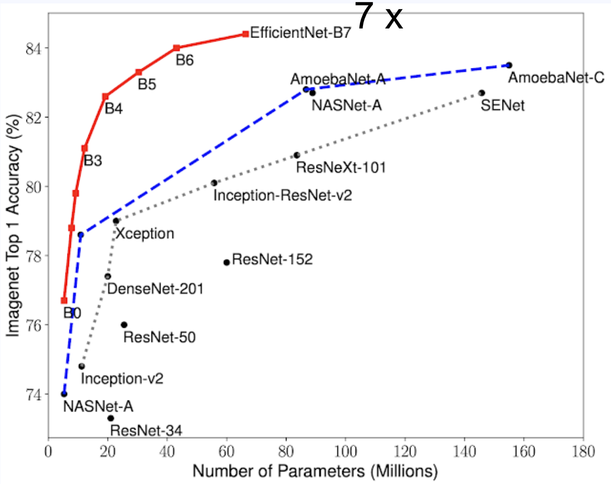
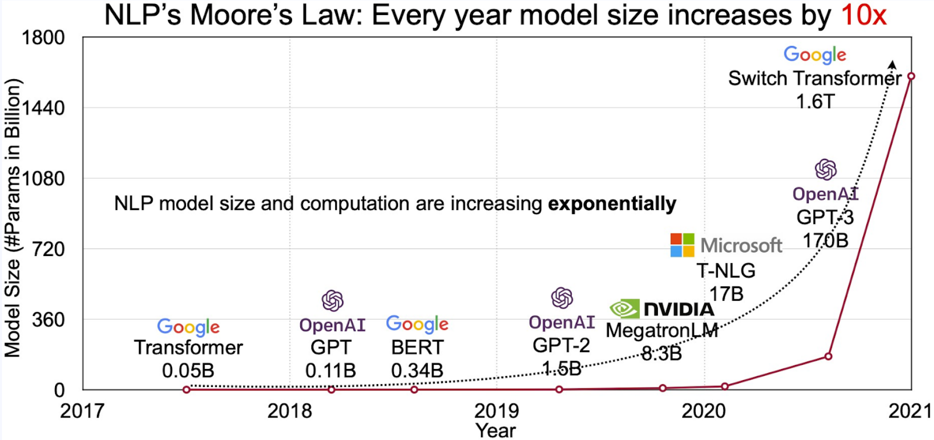


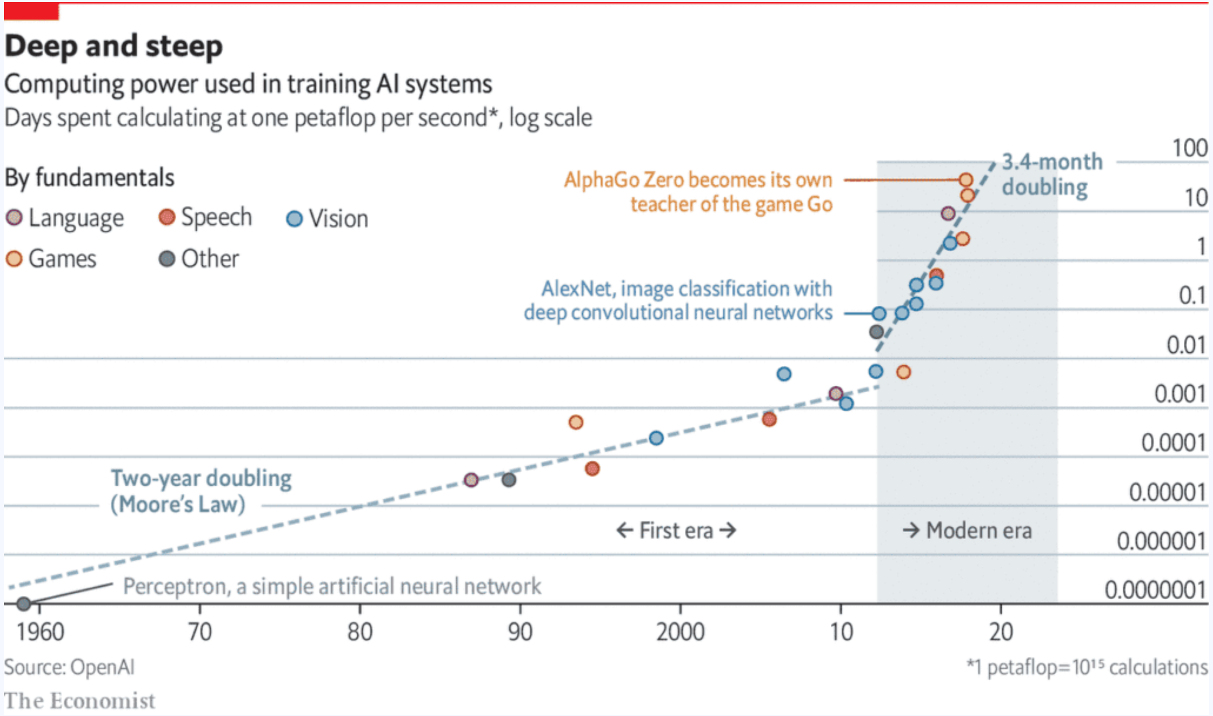
Image base



LLM

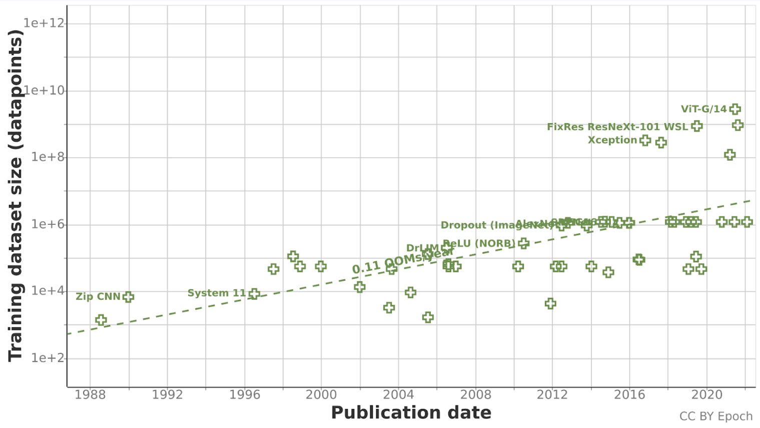
Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105-6114). PMLR.

Demand in Computer Resource

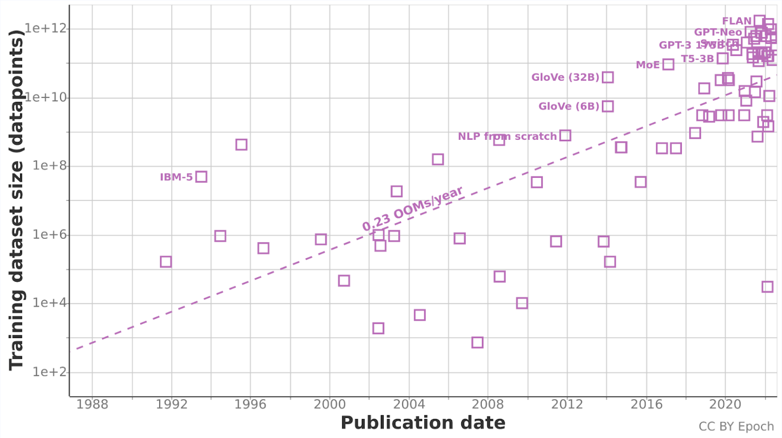


Demand in Data

Image base



NLP



arXiv:2207.02852 **[cs.LG]**

Impact of Demand

- **Socio-economic barrier**
- Infrastructure strain
- Environmental impact
- Limits in potential application

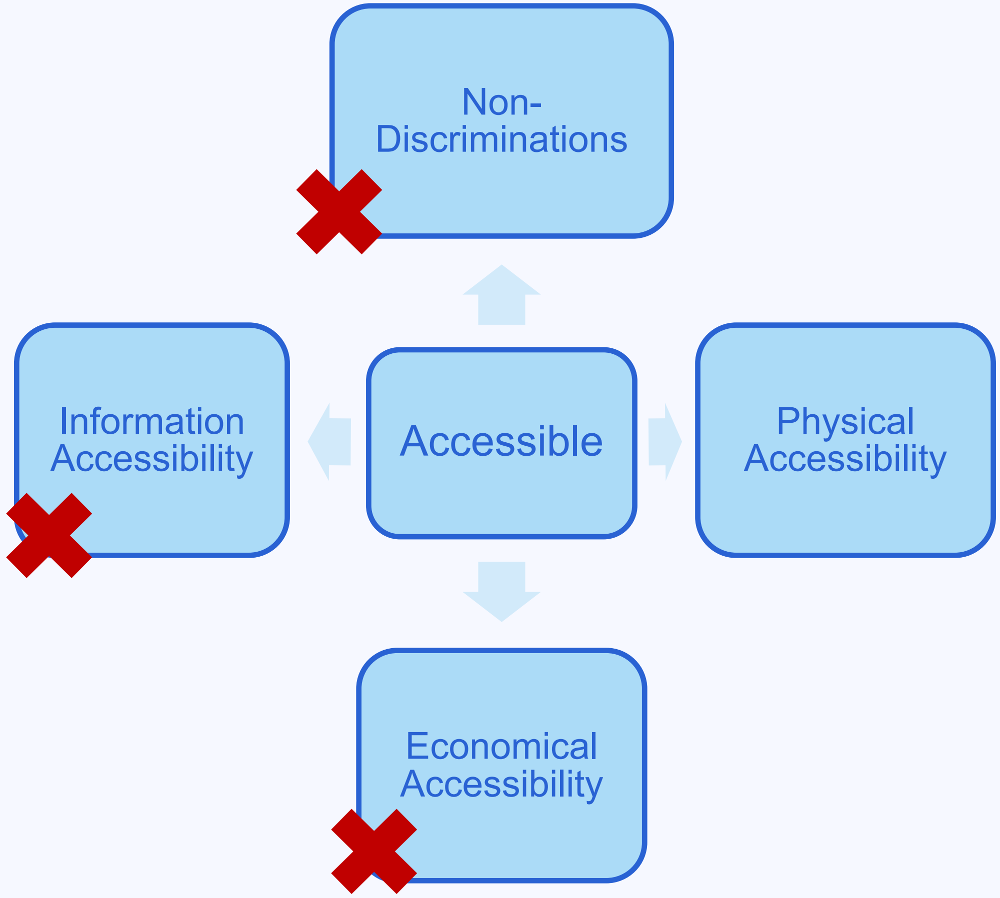
World Health Organization

Component Right to health

The WHO Constitution (1946) envisages "…**the highest attainable standard of health as a fundamental right of every human being**." Acknowledging health as a human right recognizes a legal obligation on states to ensure access to timely, acceptable, and affordable health care.

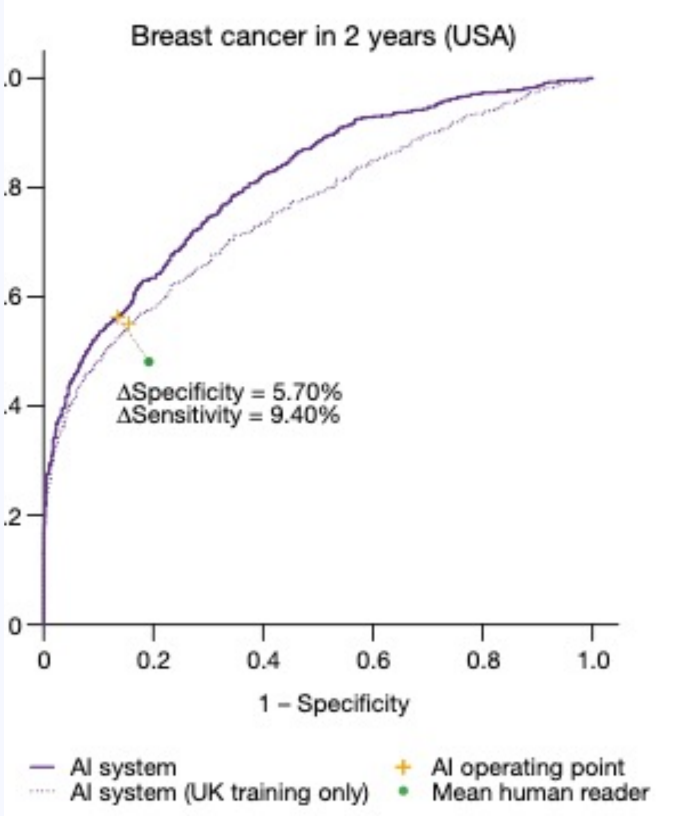
1. Availability
2. Accessibility
3. Acceptability
4. Quality

Component of in-accessibility



High Income Nation
→ Data Privacy

Middle Low-Income Nation
→ Resource Accessibility
→ Local data Availability



Effect of dataset locality

Article | Published: 01 January 2020

International evaluation of an AI system for breast cancer screening

Scott Mayer McKinney, Marcin Sieniek, [...]Shravya Shetty

Nature **577**, 89–94 (2020) | Cite this article

Local Context

- In Clinical Ai local context is crucial,
- Incident rate continuously changes
- Demographic: Population aging and mobility
- Changes in treatment planning

Local model deployment

Local training
Local validations

Estimated age-standardized incidence rates (World) in 2020, lung, both sexes, all ages

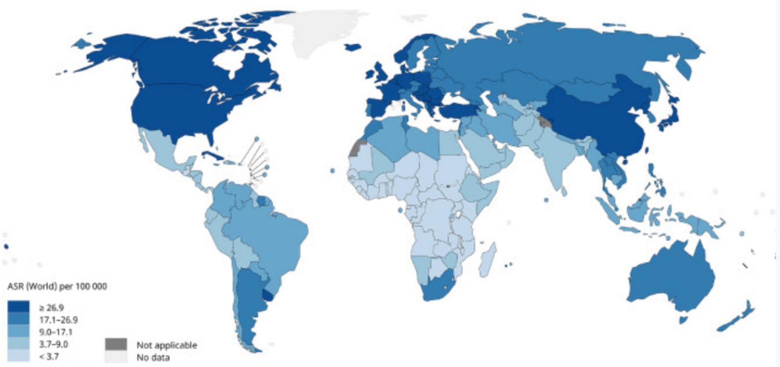
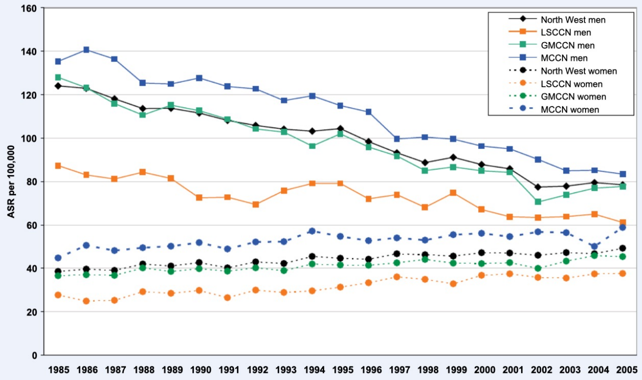


Figure 2.1: Trends in lung cancer incidence rates (ASR per 100,000) by sex and cancer network in the North West 1985 to 2005.



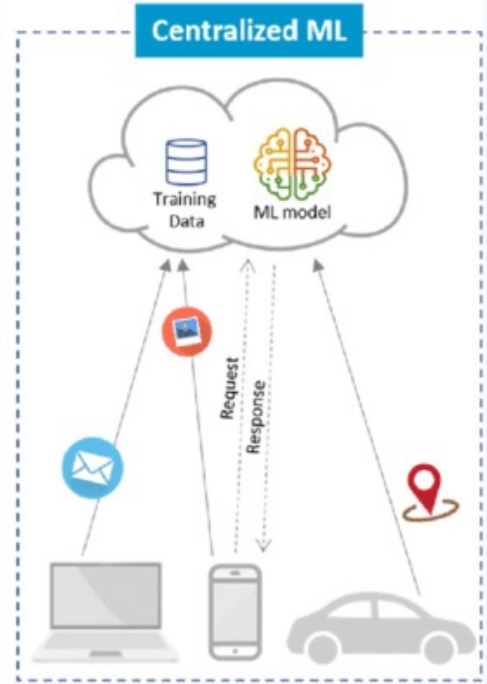
Centralize Vs Isolated Learning

Isolated Learning

- Limited Data
- Generalization Issues
- Poisson data attack

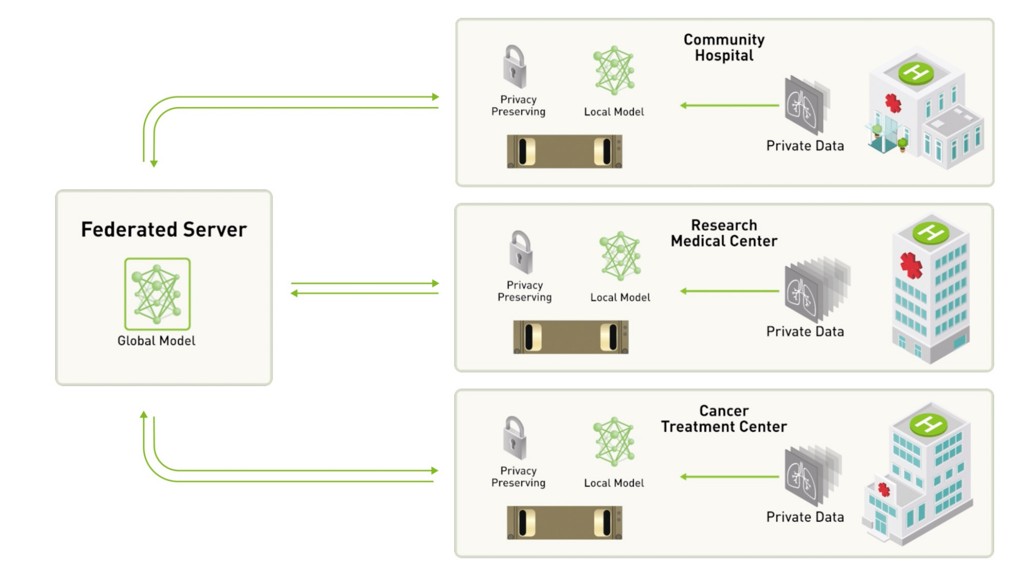
Centralized Learning

- Data privacy and Security
- Model Bias
- Network quality



Federated Learning

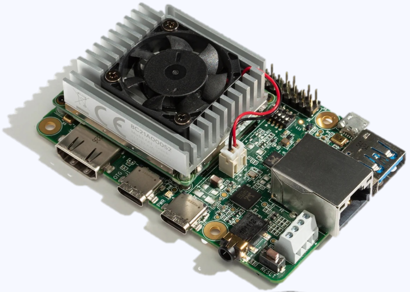
- Train and validated using local dataset
- Merge with other model for greater generalization
- No data sharing between node
- Poisson data attack are localized



Federated-AI in Malaysia

Ideally

Realistically speaking



Federated-AI in Malaysia

- Due to limit in financial resource and talent.
- Only edge-like serve can be deploy in clinic / hospital
- Introduce the restrain to what kind of ML model can be deployed.

TABLE 2. Comparison of TinyML performances against existing technologies.

Generalized

Specific

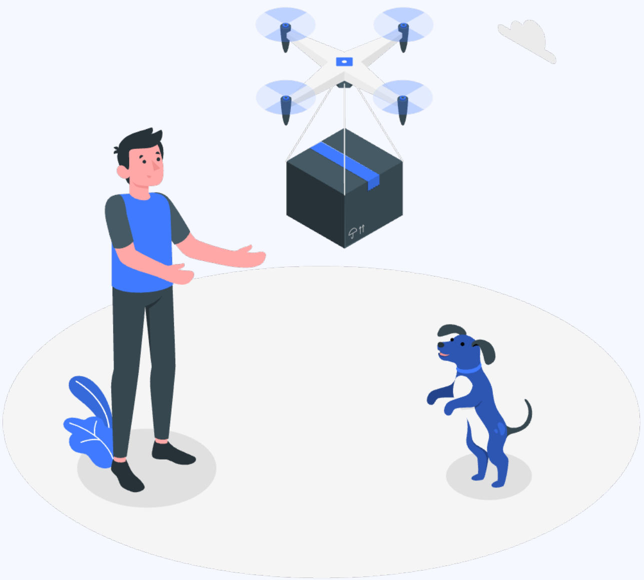
Technologies	Latency	Privacy considera-tion	Accuracy	Power Consumption	Reliability	Memory Consumption
Cloud Computing	10-500 ms	Very Low	86-94%	50-1000 W	Depends on the uninterrupted internet connectivity	GBs to TBs
Fog Computing	5-400 ms	Low	85-93%	10-100 W	Depends on active wireless connectivity	MBs to GBs
Edge computing	0.70-350 ms	Low	80-90%	1-10 W	Depends on active wireless connectivity	Few KB
TinyML	0.18-300 ms	Very High	80-90%	25-300 mW	Does not relay on network connectivity	Few KB

Abadade, Y., Temouden, A., Bamoumen, H., Benamar, N., Chtouki, Y., & Hafid, A. S. (2023). A Comprehensive Survey on TinyML. *IEEE Access*.

Tiny – ML

Subfield of AI that leverages extremely low-profile devices to execute AI algorithms, reducing the energy consumption, CO_2 emissions, and the overall cost associated with traditional AI methodologies.

ChatGPT



Tiny - ML

- Practical Applications:
  - Wearable, manufacturing sensor, health monitoring, satellite
- Low-power and Offline Capabilities:
  - Focus on microcontrollers
  - Run unplugged for extended periods
  - Remote location deployment (offline learning)
- Privacy and Ethics:
  - Processing data at the source
  - Satisfying data protection regulations :

TABLE 3. Hardware Platforms to Support TinyML

Hardware	Processor	CPU Clock	Flash Mem-ory	SRAM Size	Sensors	Power
Alif Ensemble E7	Cortex-M55 with Ethos-U55 microNPUs	400MHz	4MB	4MB	Camera and microphone	1.71-3.6V
Arduino Nicla Vision	Dual Arm Cortex M7/M4	M7: 480MHz and M4: 240MHz	2MB	1MB	Camera, microphone and IMU	3.7V Li-po battery
Infineon CY8CKIT-062S2 Pioneer Kit	Arm Cortex M4	240MHz	2MB	1MB	Accelerometer, microphone	1.8-3.3 V
Seeed Grove Vision AI Module	Himax HX6537-A	400MHz	2MB	2MB	Camera, microphone and accelerometer	5V
SiLabs Thunderboard Sense 2	Cortex-M4F	40MHz	1MB	256KB	Accelerometer and microphone	3.3V
SiLabs xG24 Dev Kit	Cortex-M33	78MHz	1.5MB	256KB	Accelerometer and microphone	3.3V
ST B-L475E-IOT01A	Arm Cortex M4	240MHz	1MB	128KB	Humidity sensor, temperature sensor, accelerometer and microphone	3-5V

Abadade, Y., Temouden, A., Bamoumen, H., Benamar, N., Chtouki, Y., & Hafid, A. S. (2023). A Comprehensive Survey on TinyML. *IEEE Access*.

Can we use Tiny-ML? No

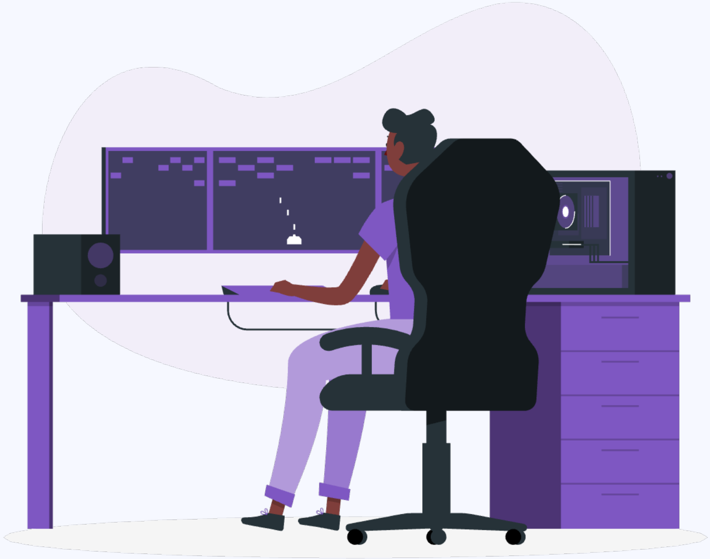
- ‘Big-Model’ is needed to process complex medical image
- Somehow mobilnet never produce good result
- Difficult task such generative and LLM are deem necessary in the future

Table 9. GPU memory and training time on ImageNet; memory indicates that per GPU and time indicates that per iteration.

Architecture	Top-1 (%)	memory (MB)	time (s)
ResNet50 [14]	24.01	3929	0.31
ResNet152 [14]	22.16	7095	0.63
DenseNet264 [18]	22.15	9981	0.60
DSNet50	22.49	4777	0.37
DS2Net50	22.03	5133	0.39

Backbone Model	**Transfer Learning Approach** [2]	**Mean Training Accuracy**	**Mean Test AUC**
	Initiating transfer learning		
ResNet50	IR	**0.935**	0.796
ResNet50	CR	0.879	**0.831**
ResNet50	CI	0.868	**0.827**
ResNet50	XR	0.819	0.806
ResNet50	XI	0.852	**0.831**
DenseNet121	IR	0.916	0.803
DenseNet121	CR	0.807	0.800
DenseNet121	CI	0.784	0.779
enseNet121	XR	0.799	0.781
enseNet121	XI	0.864	0.826
	Concatenating transfer learning		
ResNet50	I + C	**0.935**	0.780
ResNet50	I + X	0.930	0.776
enseNet121	I + C	**0.935**	0.802
enseNet121	I + X	0.914	0.813
	Co-training transfer learning		
ResNet50	C ∪ X	0.855	0.790
enseNet121	C ∪ X	0.775	0.826

Huang, G. H., Fu, Q. J., Gu, M. Z., Lu, N. H., Liu, K. Y., & Chen, T. B. (2022). Deep transfer learning for the multilabel classification of chest X-ray images. *Diagnostics*, *12*(6), 1457

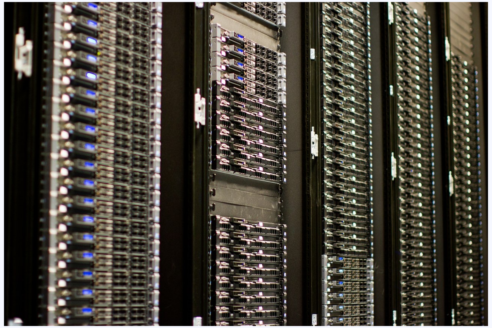


02

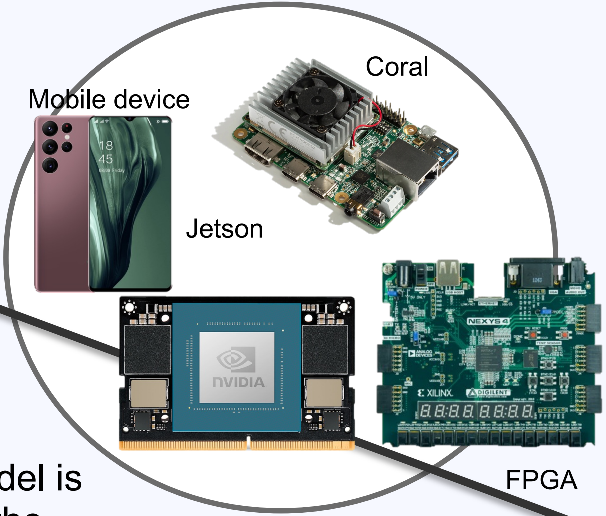
What is Lite-Ai

Lite-Ai

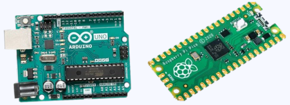
is a middle alternative



Big – ML Optimize for big server



- Which ML model is optimized for the middle ?
- We have the device but not the ML model



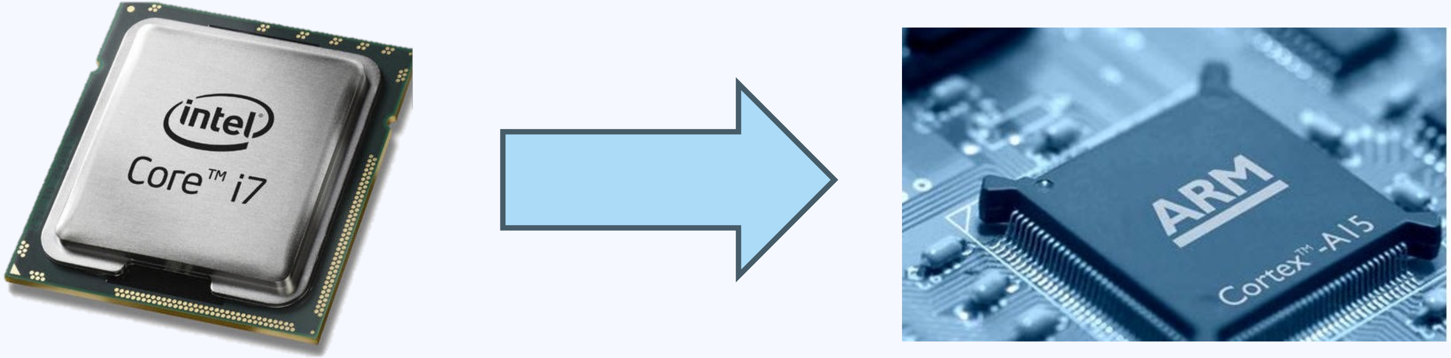
Tiny - ML optimize for microcontroller

Lite – Ai

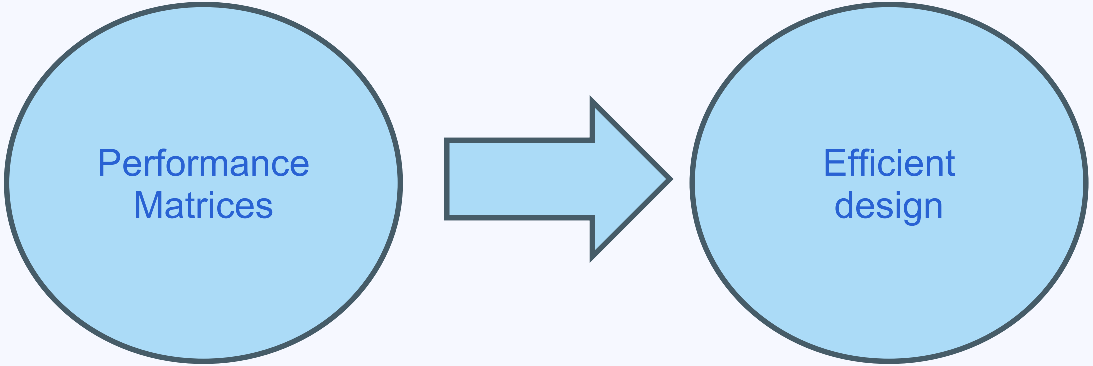
- Not attempt to shrink or minimize the complexity!
- Is attempt to **repackage** or **re-optimize** existing Big AI Model to be more efficient
- Make it **'Lite'** enough to run on middle-size computer (~Fog computing)
    - RAM – 1 – 8 Gig
    - GPU is not essential
- To be execute in environment and infrastructure that have **'adequate-resource'**
    - Continuous power
    - Stable connectivity (not fast)
    - No aircon but climate fluctuations are negligible

The Tomark Aero Viper SD4
Even though it is a 'Lite' aircraft but still highly complex

Transitioning



Paradigm Shift

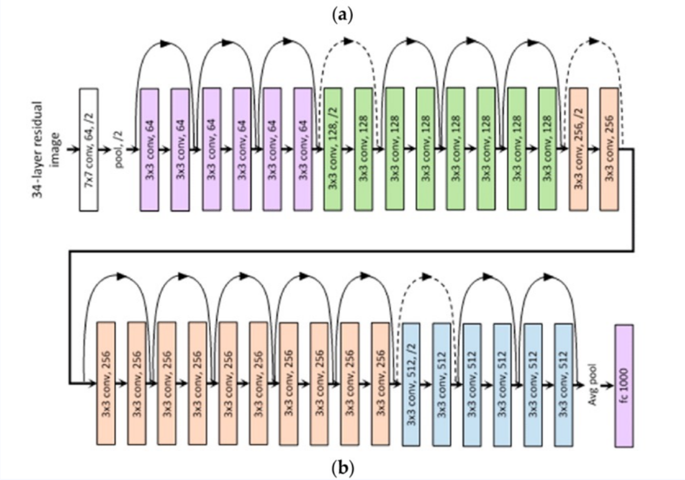


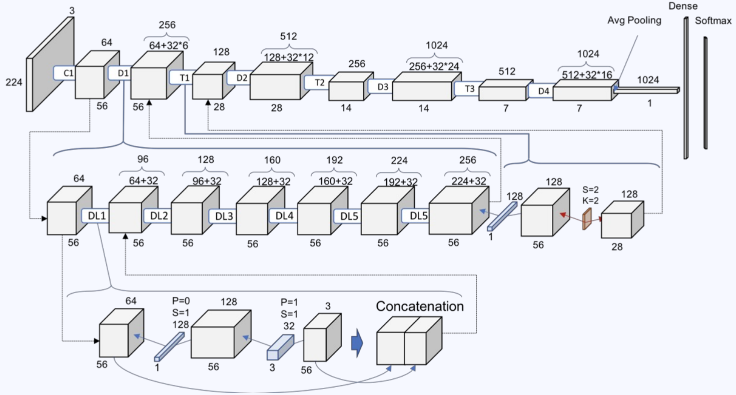
03

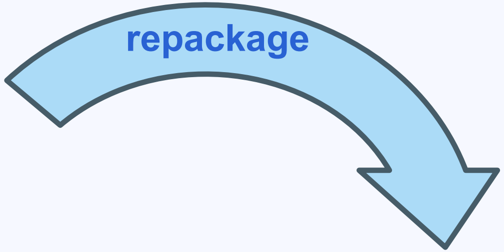
HOW ?

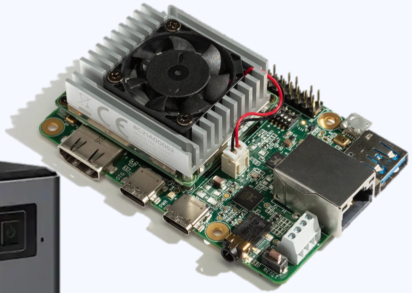
Is it even possible

still work in **progress**


(a)
34-layer residual
image
7x7 conv, 64, /2
pool, /2
3x3 conv, 64
3x3 conv, 128, /2
3x3 conv, 128
3x3 conv, 256, /2
3x3 conv, 256
3x3 conv, 512, /2
3x3 conv, 512
Avg pool
fc 1000
(b)


Dense
Avg Pooling
Softmax
C1
D1
T1
D2
T2
D3
T3
D4
DL1
DL2
DL3
DL4
DL5
Concatenation
64+32*6
128+32*12
256+32*24
512+32*16
S=2
K=2
P=0
S=1
P=1


repackage




repackage

Example of Method

On-the-Fly Loading

Loading dataset and per-processing mapping is done per-batch

Anomalous ROI

Extracting saliency mapping before training

Model Repacking

Repacking and re-modelling the model

Clinical Data Size

Cloud Healthcare API > Documentation > Resources

Was this helpful?

NIH Chest X-ray dataset

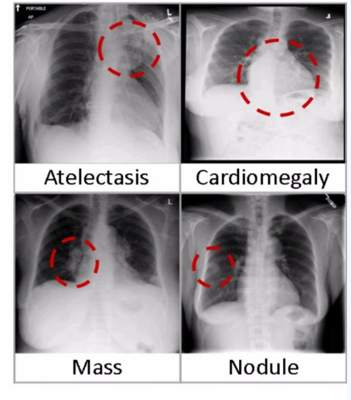
Send feedback

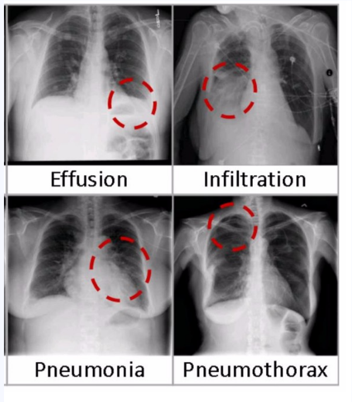
The NIH Chest X-ray dataset consists of 100,000 de-identified images of chest x-rays. The images are in PNG format.

The data is provided by the NIH Clinical Center and is available through the NIH download site: https://nihcc.app.box.com/v/ChestXray-NIHCC

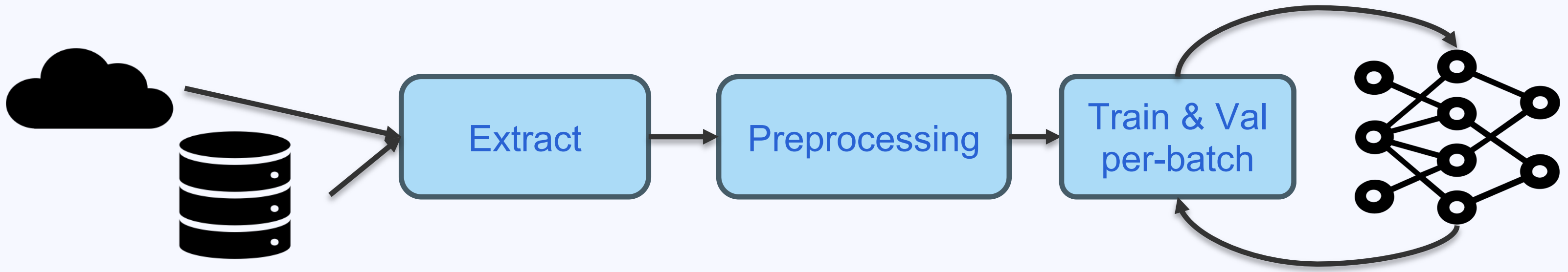
You can also access the data via Google Cloud, as described in Google Cloud data access.

- Total size of sample is 43 Gig
- Impossible to load ALL in RAM

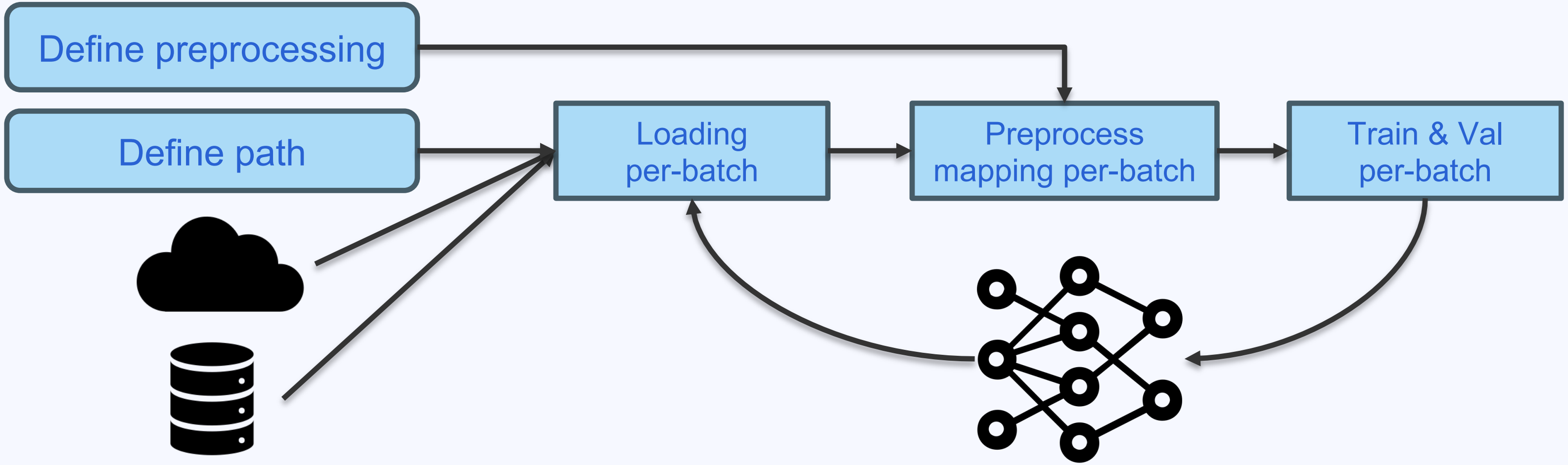




Full data loading



On-the-fly


Define preprocessing
Define path
Loading per-batch
Preprocess mapping per-batch
Train & Val per-batch

TF. Example

python Copy code

```python
# Import necessary libraries
import tensorflow as tf
from tensorflow.keras import layers
from tensorflow.keras.models import Sequential

# Define the dataset
train_ds = tf.keras.utils.image_dataset_from_directory(
    path_directory,
    batch_size=batch_size)

# Define a function for image preprocessing
def process_images(image, label):
    image = tf.image.per_image_standardization(image)
    return image, label

# Map the function across the dataset
train_ds = train_ds.map(process_images)

# Define a model for binary classification
model = # define model

# Compile the model
model.compile(optimizer='adam',
              loss=tf.keras.losses.BinaryCrossentropy(),
              metrics=['accuracy'])

# Fit the model
model.fit(train_ds, epochs=10)
```

Dataset obj

Preprocess func()

Attach map func()

Give data obj

- For advance input and output require tf.data
- Sorry not familiar with pytorch

Pro & Con of 

- Advantageous:
    - Allowing bigger and complex model to be load into the RAM
    - Adding room to increase input data dimension ← more information
    - Take advantage of parallel processing, CPU threads handling data loading and preprocessing while GPU threads handle model training

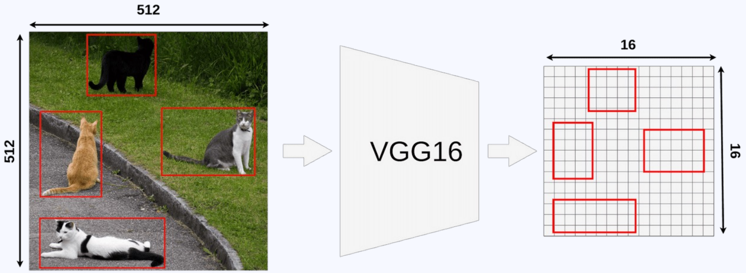
Pro & Con of

- Disadvantageous:
    - Complex input (multi-type input) with complex target (regression, multilabel, multitype) will be very difficult to set-up
    - Slow data extraction (pulling from cloud) can cause GPU to wait for data input; therefore, the I/O speed matter

Anomalous ROI

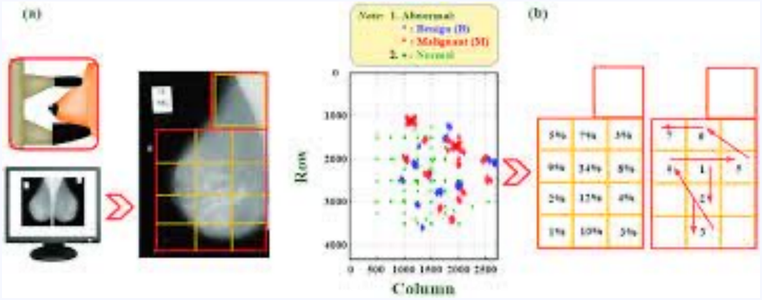
Region of Interest (ROI) extraction

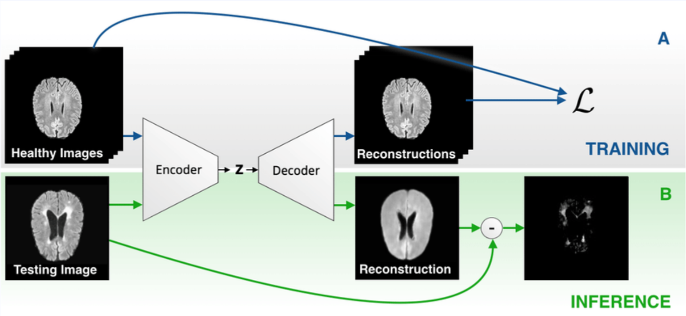
- Extract ROI from image
- Region proposal network
- Reduces computation resource improving efficiency and speed, as the network only needs to focus on relevant regions



Anomalous ROI Extraction

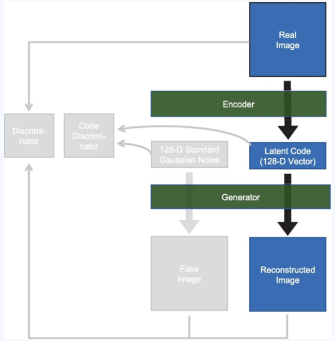
- Anomalous region extraction focuses on identifying regions in the image that significantly deviate from the norm.
- This is primarily used in anomaly detection tasks,





Autoencoder

Baur, C., Denner, S., Wiestler, B., Navab, N., & Albarqouni, S. (2021). Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study. *Medical Image Analysis*, *69*, 101952.



GAN

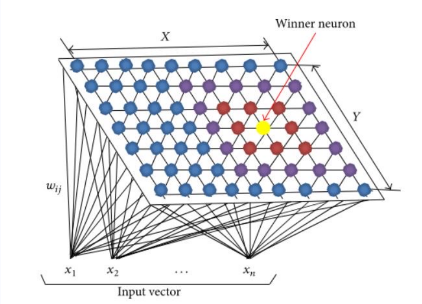
Nakao, T., Hanaoka, S., Nomura, Y., Murata, M., Takenaga, T., Miki, S., ... & Abe, O. (2021). Unsupervised deep anomaly detection in chest radiographs. *Journal of Digital Imaging*, *34*, 418-427.

senafas x2

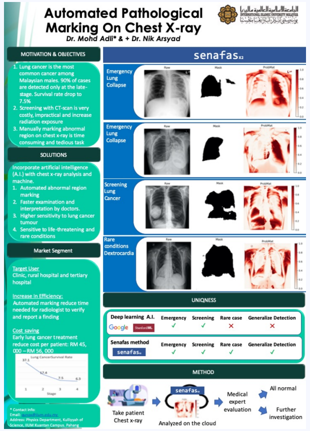
Data: Batch of images
Result: 2D Pixel Intensity Distribution (probMat)
initialization;
probMat = [ ][ ];
for *x in image width, y in image height*, **do**
  pxiLs = [pixel intensity at (x,y) for image in image batch];
  probFunc( ) = non-parametric probability function of pxiLs;
  probMat[x][y] = [probFunc(i) for i in range 0 to 255];
end

Fig. 1: The pseudocode for producing the ProbMat.

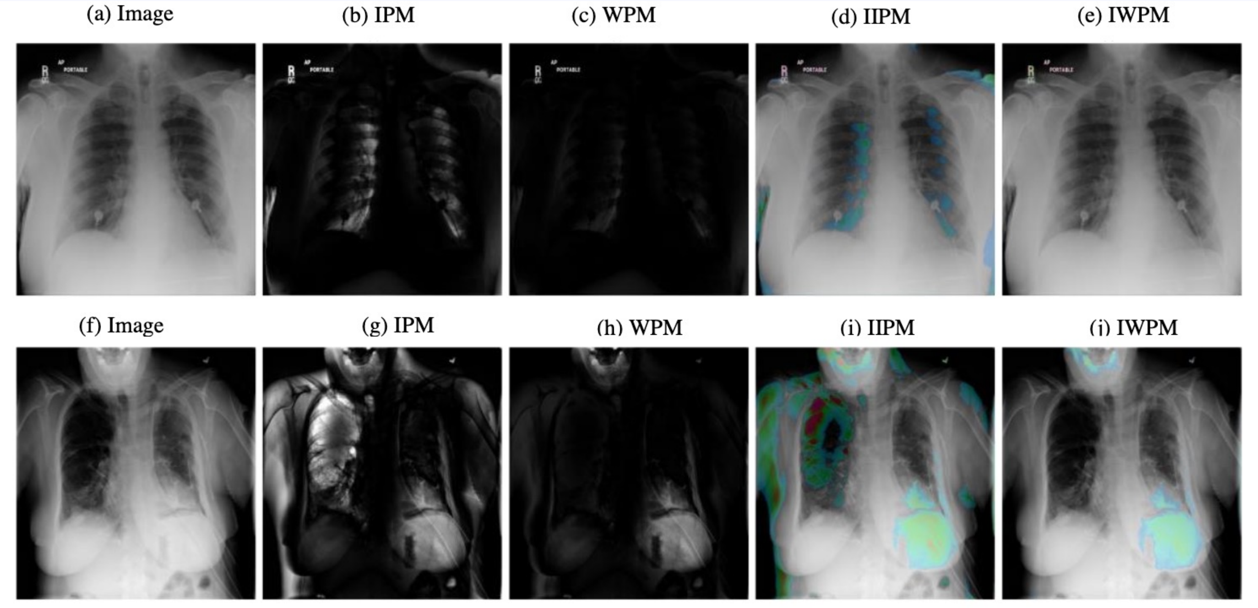
ProbMat
+
SOM



senafas X2



Allow image resolution to be reduce while still 'noting' the CNN-model there's an anomaly in this region



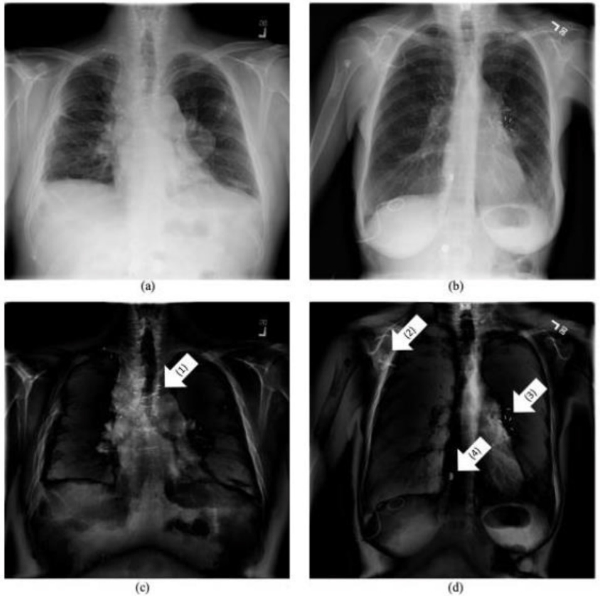
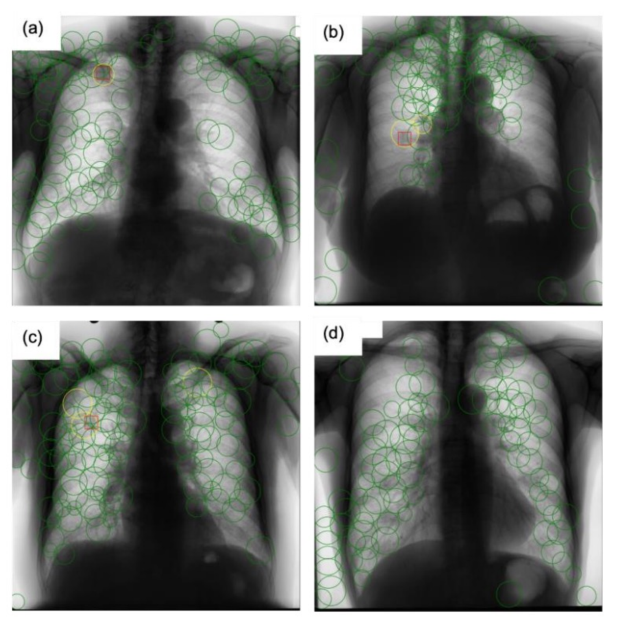


Fig. 6: Chest radiograph with foreign body, (a) and (b).
The resulting WPM images (c) and (d) clearly show the foreign bodies, arrow (1)-(4).

Hough Circle Transformation



Model Repacking

- Does not reduce the baseline memory much, but immensely reduce training memory

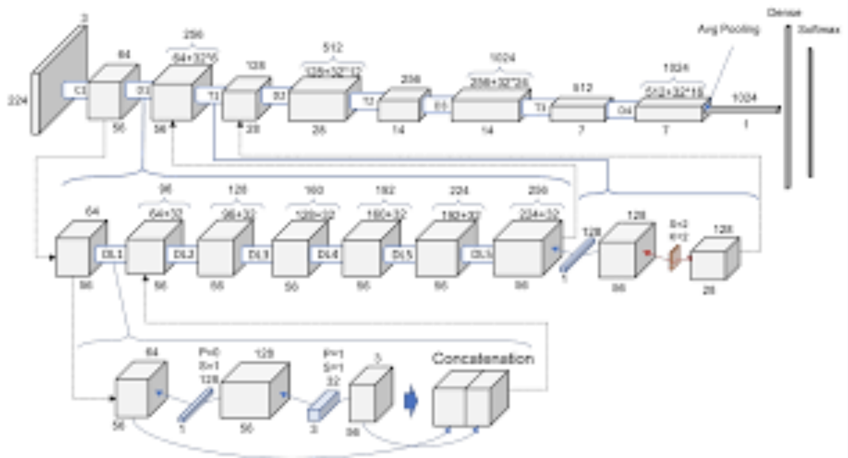
- Require careful understanding the function of each layer, the activation function and optimization
- Require dedicated MSc / PhD

- Example of Model repacking
  - Easy : Reducing channel from 3 to 1
  - Intermediate: Combining residual layer
  - Hard: ”densing” a model

- Cheat technique is to use knowledge distillation technique.

Example of

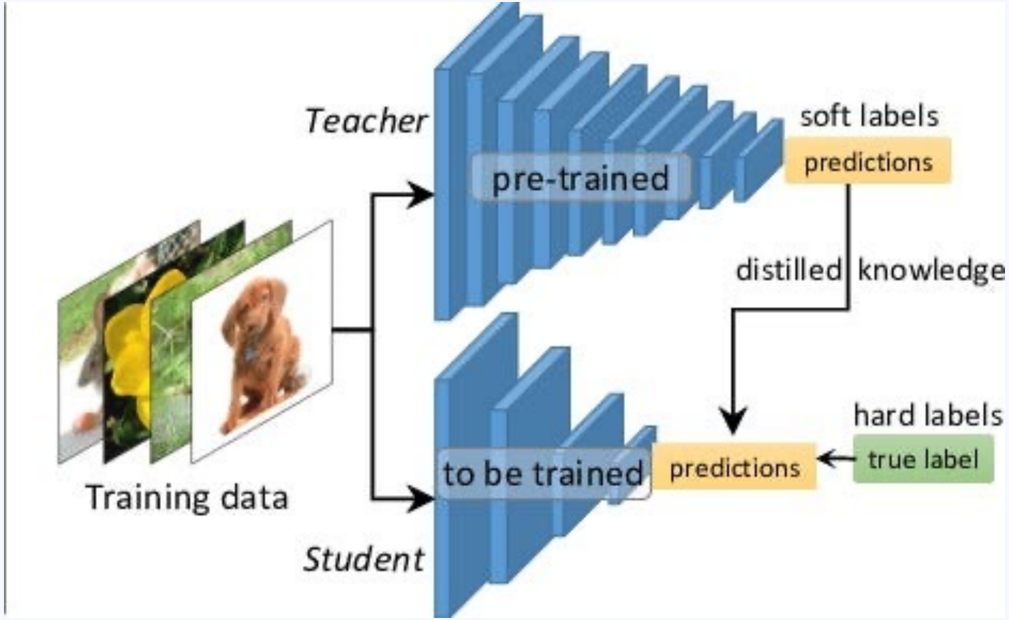
- A model inspire by DenseNet121
- Can take 512x512 image as input
- Use 1 channel instead of 3
- Aggressive feature reduction



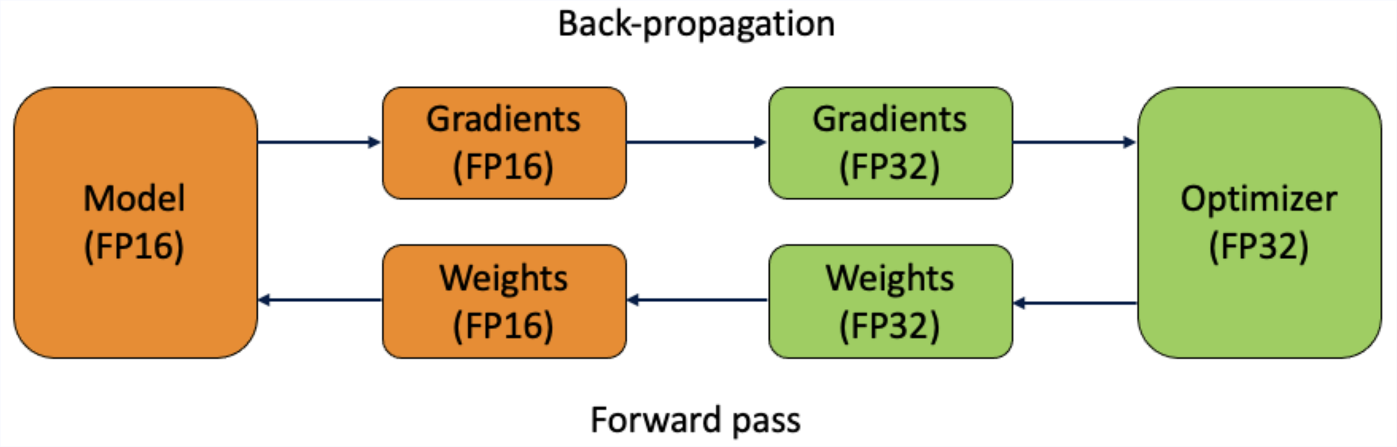
python　Copy code

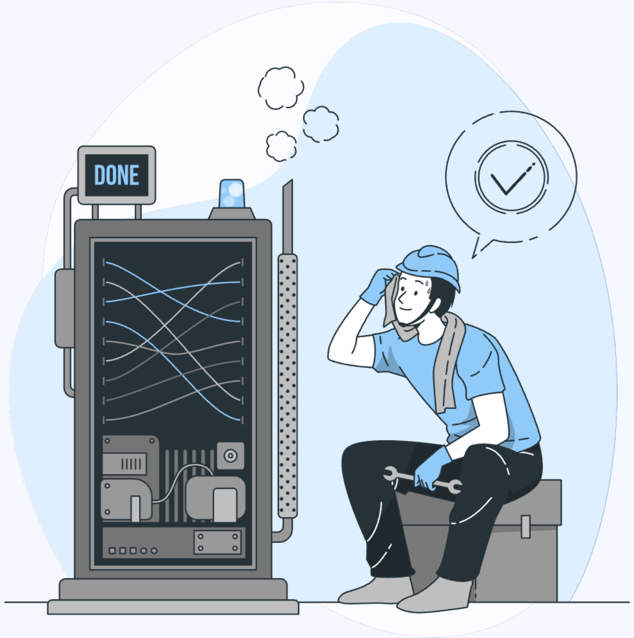
```python
from keras.models import Model
from keras.layers import Conv2D, AveragePooling2D, Flatten, Dense, Activati
from keras.regularizers import l2
input_img = Input(shape=(512, 512, 1))
x = Conv2D(32, (3, 3), padding='same', kernel_regularizer=l2(0.01))(input_i
x = Activation('swish')(x)
x = AveragePooling2D(pool_size=(2, 2))(x)
x = Conv2D(64, (3, 3), padding='same', kernel_regularizer=l2(0.01))(x)
x = Activation('swish')(x)
x = AveragePooling2D(pool_size=(2, 2))(x)
x = Conv2D(128, (3, 3), padding='same', kernel_regularizer=l2(0.01))(x)
x = Activation('swish')(x)
x = AveragePooling2D(pool_size=(2, 2))(x)
x = Flatten()(x)
x = Dense(256, kernel_regularizer=l2(0.01))(x)
x = Activation('swish')(x)
res = x
x = Dense(128, kernel_regularizer=l2(0.01))(x)
x = Activation('swish')(x)
x = Dense(256, kernel_regularizer=l2(0.01))(x)
x = Add()([x, res])
x = Activation('swish')(x)
x = Dense(5)(x)
out = Activation('softmax')(x)
model = Model(inputs=input_img, outputs=out)
model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['
```

Knowledge Distillation method



Mixed Precision





04

Conclusion
Q & A

Conclusion

- Due limit is resource while at the same time requiring complex AI-model. A new type of AI-Model is needed
- Lite-AI attempt to **repackage** or **re-optimize** existing Big AI Model to be more efficient
- It also develop to increase Clinical Ai accessibility to more people

In research and development, we need to strongly consider its **accessability** to people of less fortunate

Q & A

Storyset

Create your Story with our illustrated concepts. Choose the style you like the most, edit its colors, pick the background and layers you want to show and bring them to life with the animator panel! It will boost your presentation. Check out **how it works**.

Pana

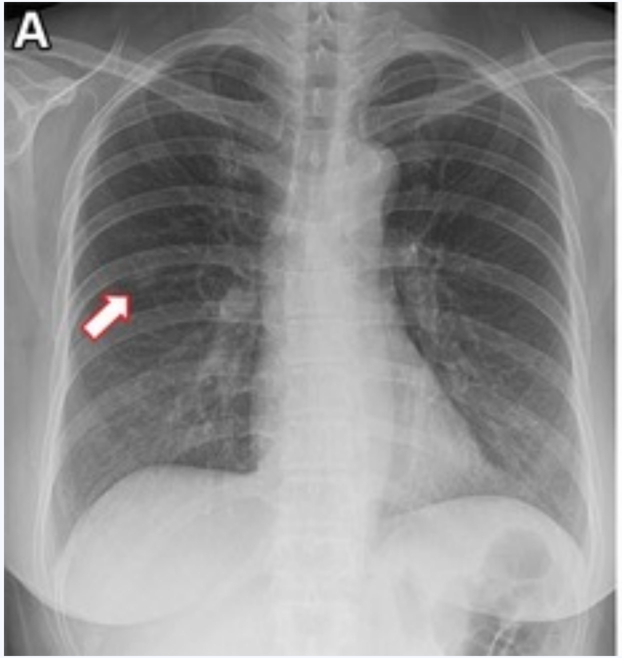
Amico

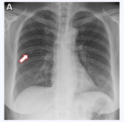
Bro

Rafiki

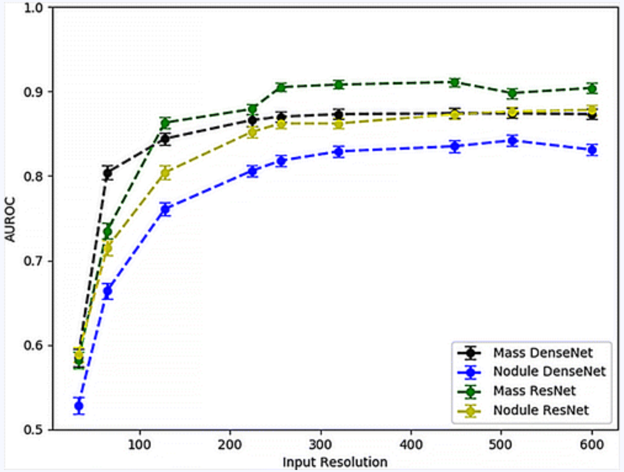
Cuate

Reduce input size? ~No





256 x 256



Sabottke, C. F., & Spieler, B. M. (2020). The effect of image resolution on deep learning in radiography. *Radiology: Artificial Intelligence*, *2*(1), e190015.