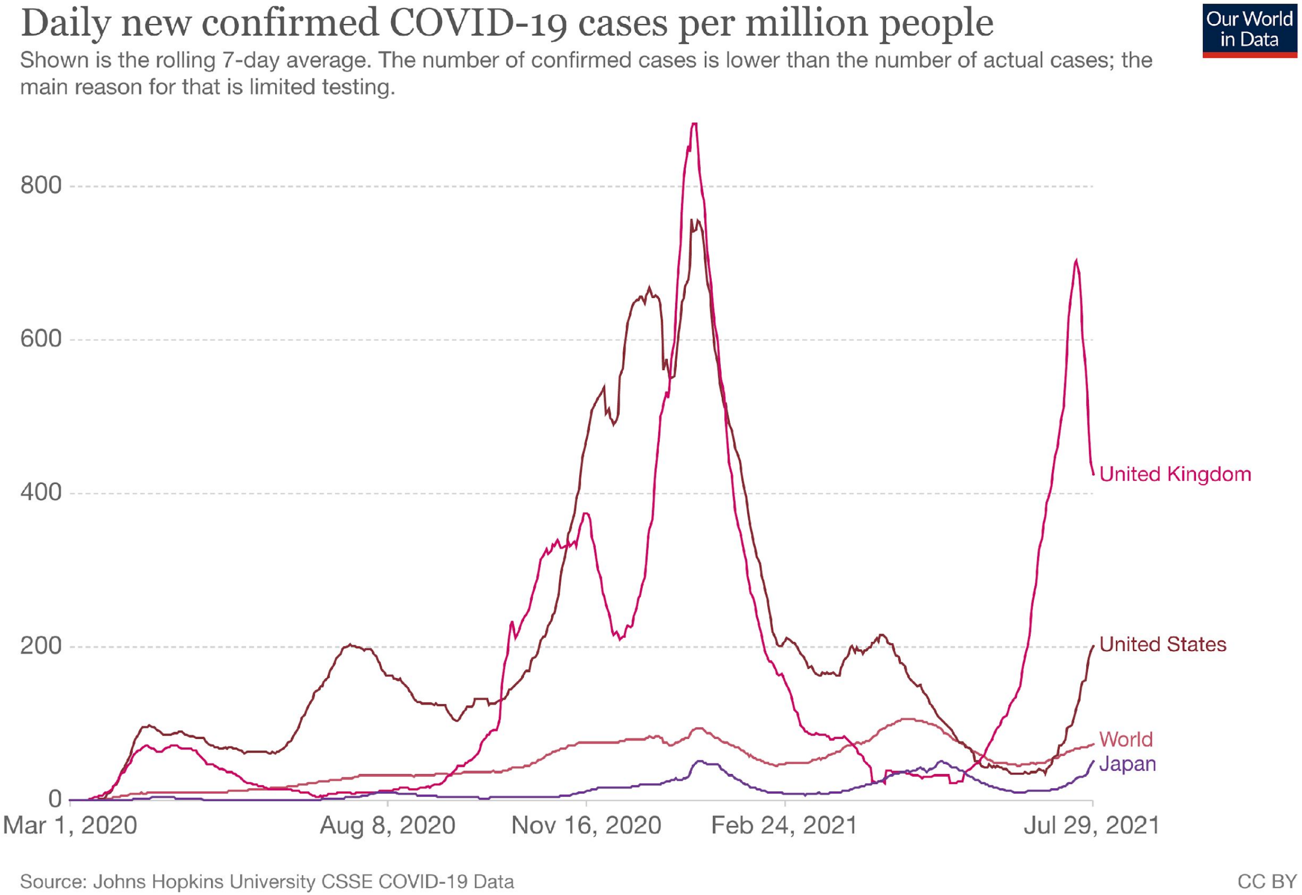
INTERNATIONAL CONFERENCE ON PARALLEL PROCESSING
ICPP / 2021 / CHICAGO / USA
acm In-Cooperation
sighpc
AUGUST 9-12, 2021

COMPUTECOVID19+

Accelerating COVID-19 Diagnosis and Monitoring via High-Performance Deep Learning on CT Images

Garvit Goel*, Atharva Gondhalekar*, Jinyuan Qi*,
Zhicheng Zhang†, Guohua Cao*, Wu Feng*

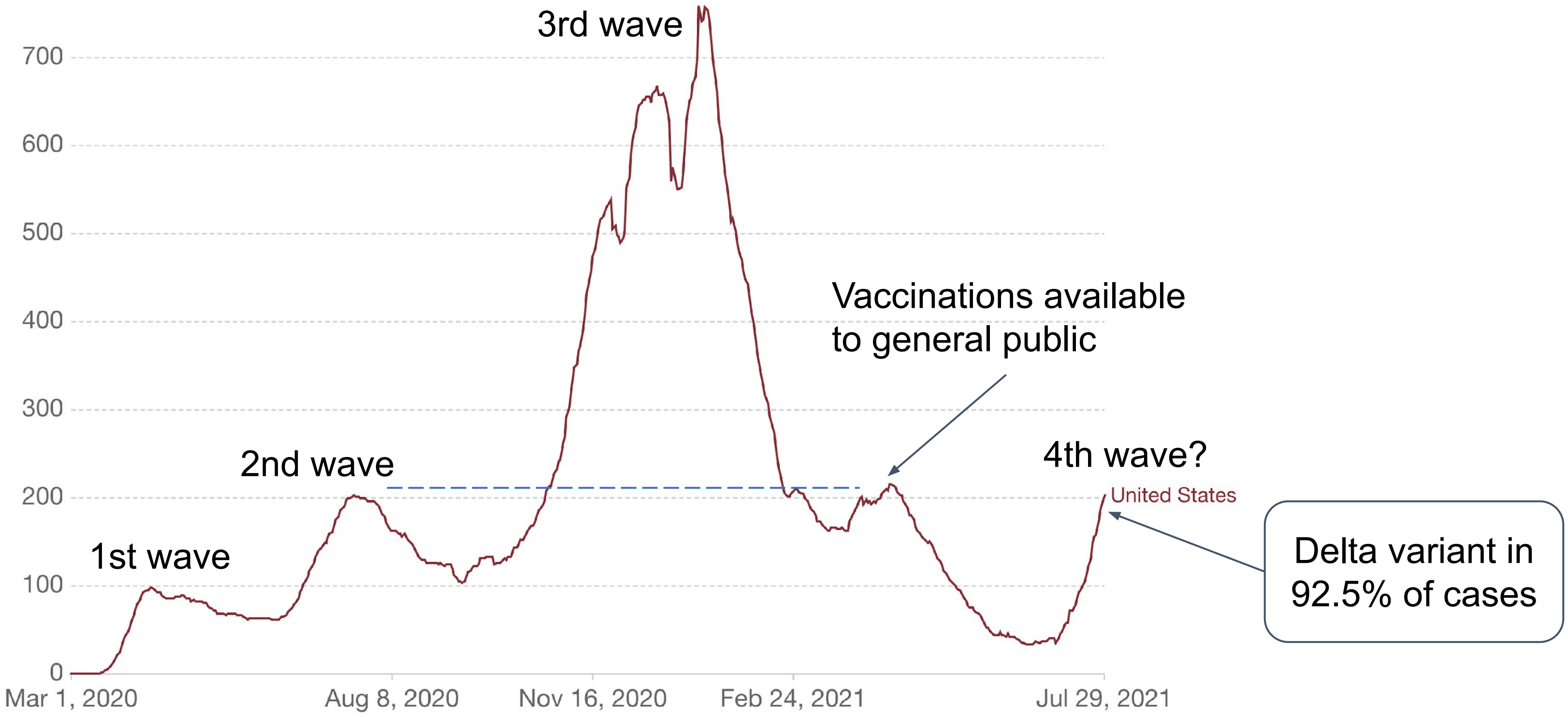
*Virginia Tech and †Stanford University



Daily new confirmed COVID-19 cases per million people
Shown is the rolling 7-day average. The number of confirmed cases is lower than the number of actual cases; the main reason for that is limited testing.
Our World in Data
800
600
400
200
0
Mar 1, 2020
Aug 8, 2020
Nov 16, 2020
Feb 24, 2021
Jul 29, 2021
United Kingdom
United States
World
Japan
Source: Johns Hopkins University CSSE COVID-19 Data
CC BY

Daily new confirmed COVID-19 cases per million people

Shown is the rolling 7-day average. The number of confirmed cases is lower than the number of actual cases; the main reason for that is limited testing.



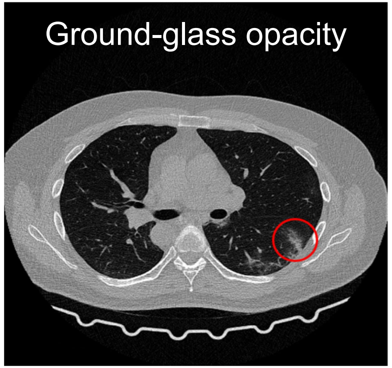
Source: Johns Hopkins University CSSE COVID-19 Data

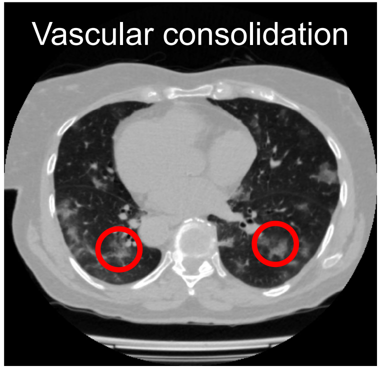
RT-PCR

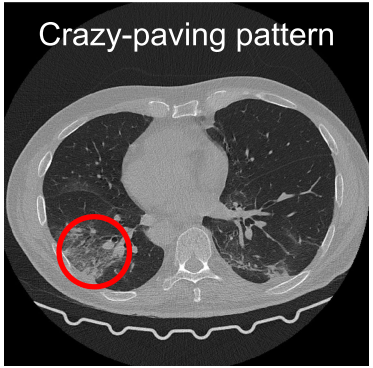
	RT-PCR ("The Gold Standard")
Principle	Analysis of RNA/DNA
Accuracy	**67% accuracy** (end-to-end)* Source: Johns Hopkins University, June 2020
Turnaround Time	**~240 minutes (4 hours)** per test … with multi-day turnaround time

RT-PCR vs. ComputeCOVID19+

	RT-PCR (“The Gold Standard”)	*ComputeCOVID19+*
Principle	Analysis of RNA/DNA	Chest CT scan
Accuracy	**67% accuracy** (end-to-end)* Source: Johns Hopkins University, June 2020	**91% accuracy**
Turnaround Time	**~240 minutes (4 hours)** per test … with multi-day turnaround time	**< 1 minute** per test … with 5-minute turnaround time



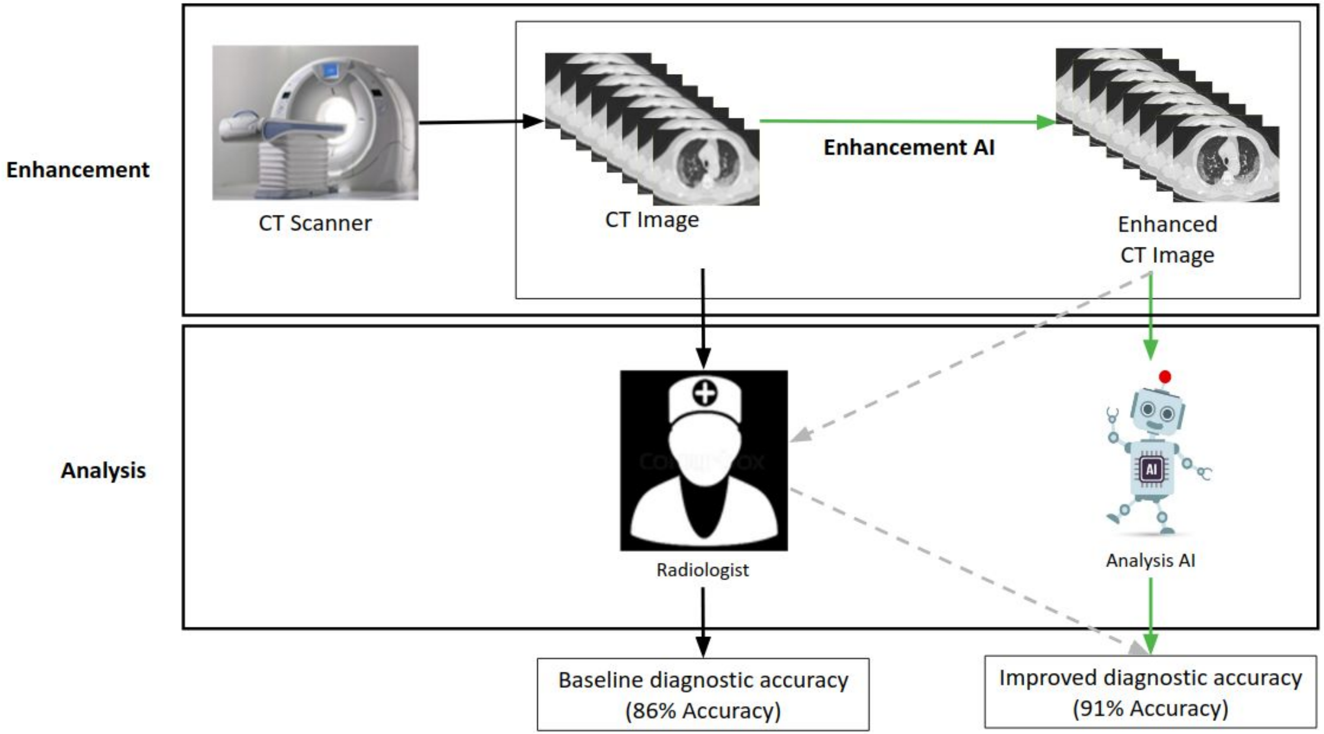




Contributions of our ComputeCOVID19+ Framework

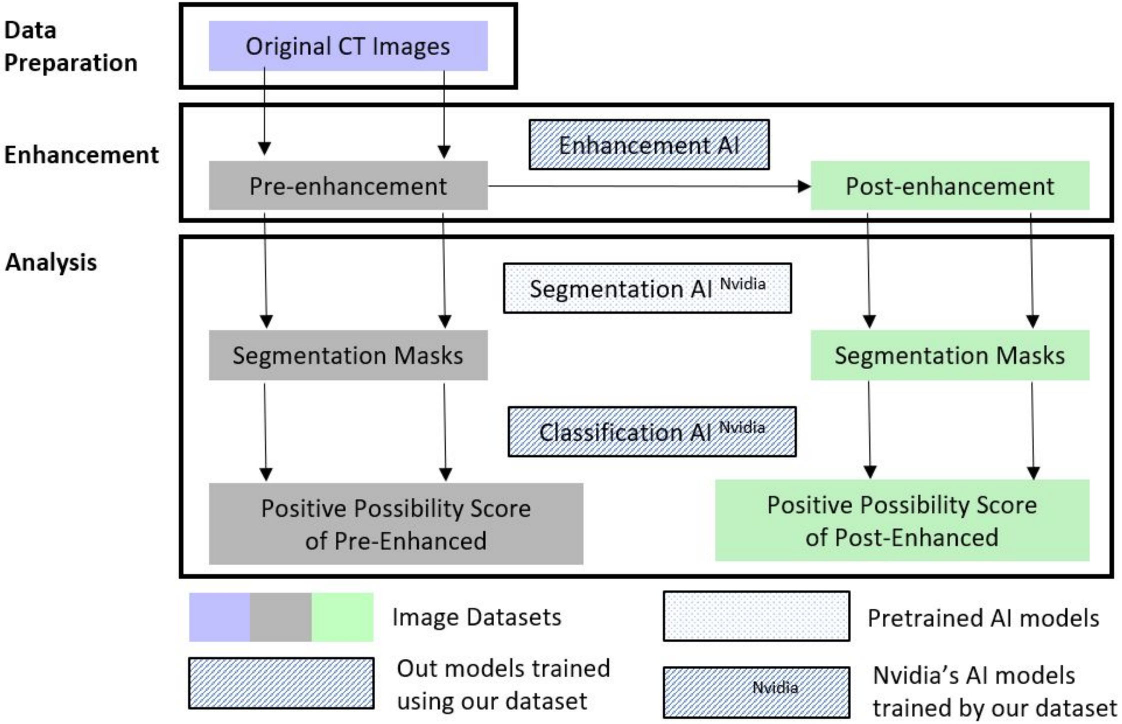
- Novel algorithms and software for high-precision CT image construction and interpretation of COVID-19
- Performance evaluation with respect to speed and accuracy
- Validation of ComputeCOVID19+ with clinical COVID-19 data
- Broader applicability of ComputeCOVID19+ to biomedical imaging

ComputeCOVID19+ Framework

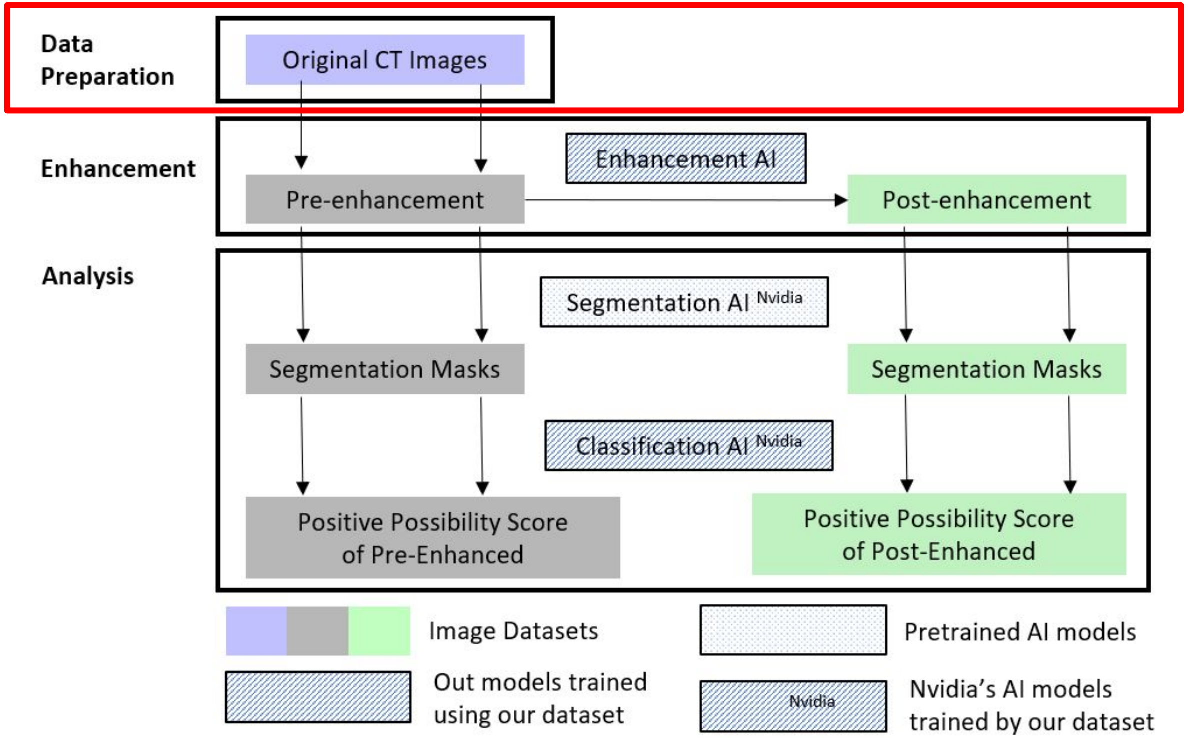


Green arrows show the improved workflow with ComputeCOVID19+

Workflow for *ComputeCOVID19+*



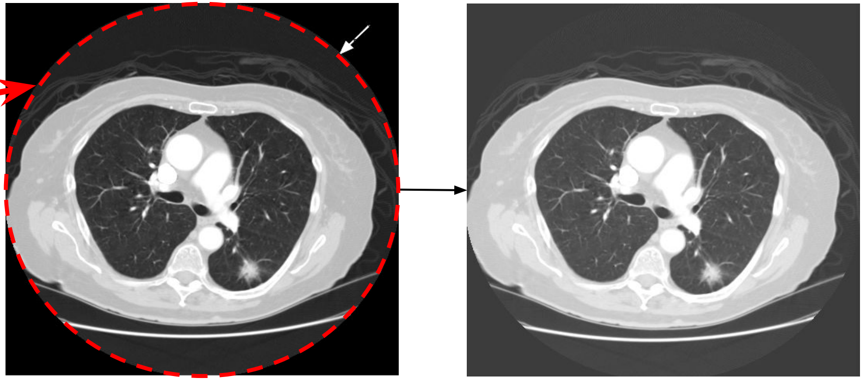
Workflow for *ComputeCOVID19+*



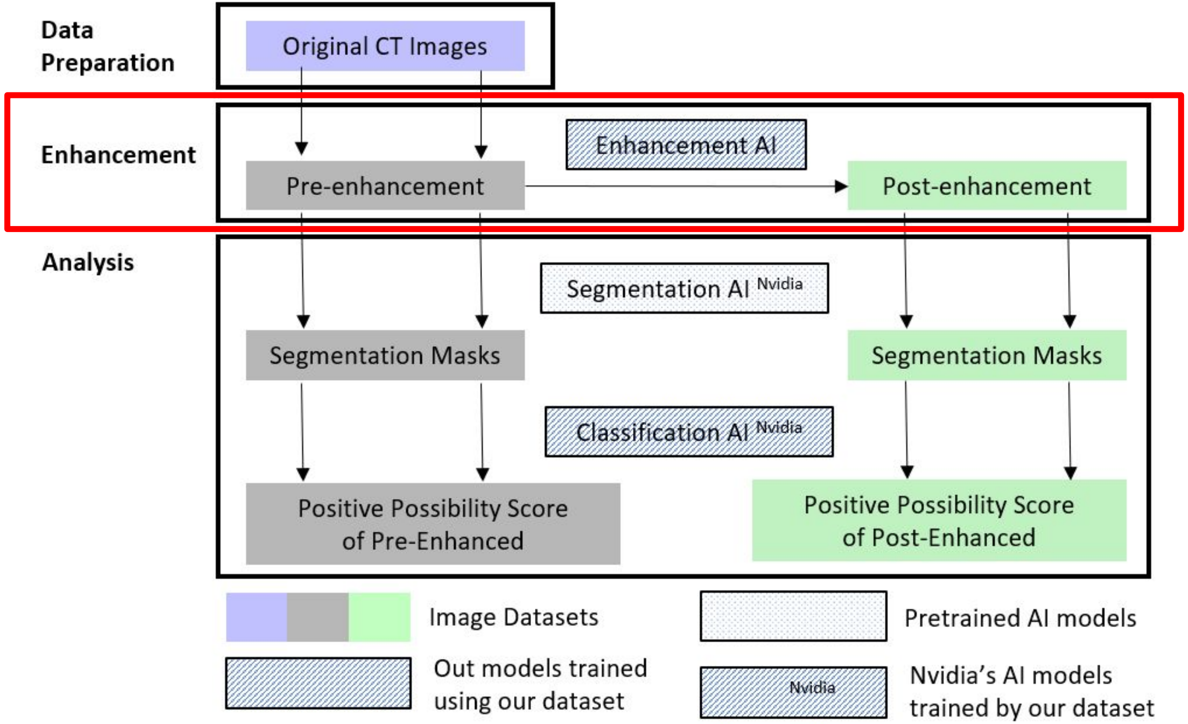
Data Preparation

Data Source	Description
Mayo Clinic	Eight healthy chest CT scans & associated projection data at full & quarter dosage
Medical Imaging Databank of Valencia Region (BIMCV)	X-ray & CT scans of 34 COVID-19 patients
Medical Imaging and Data Resource Center (MIDRC)	229 CT scans of COVID-19 patients
Lung Image Database Consortium (LIDC)	1301 healthy chest CT scans

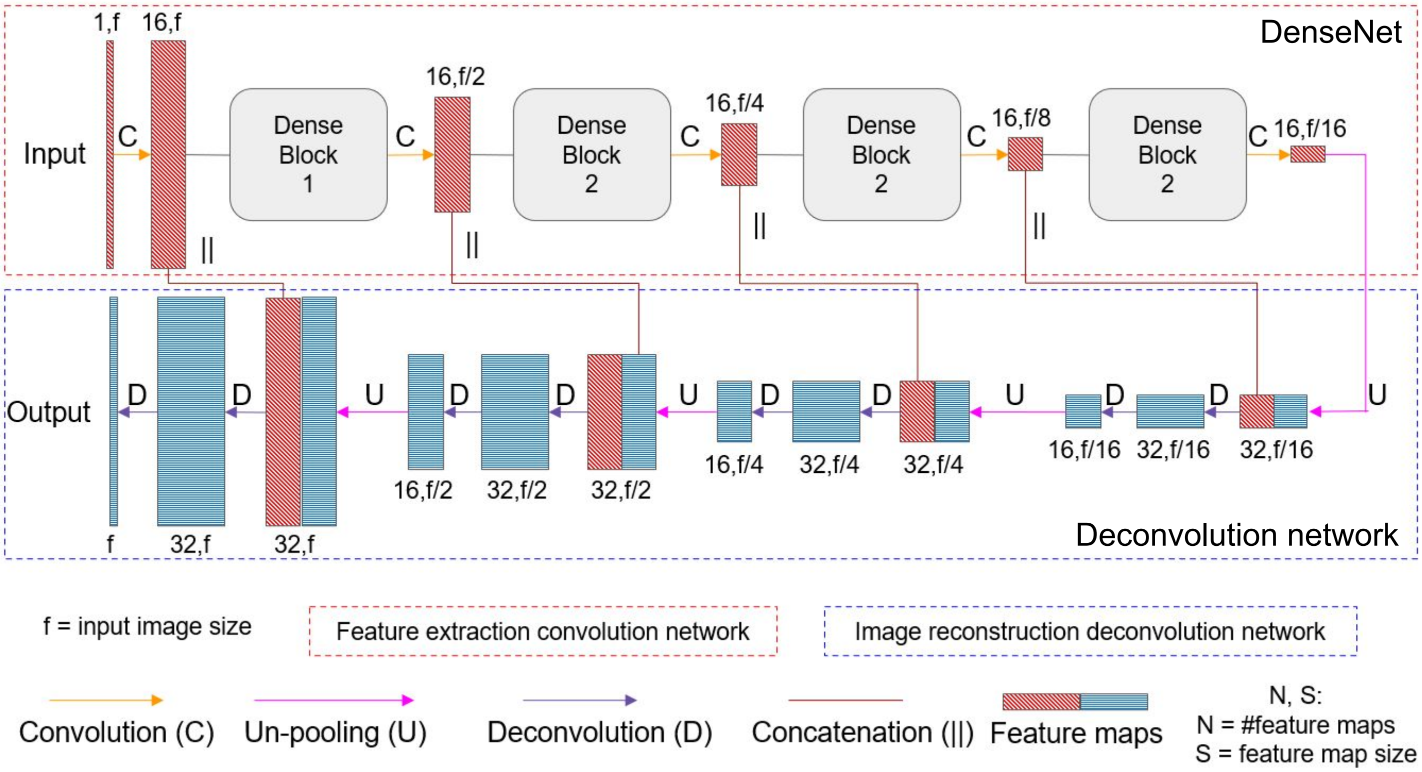
Processing for data consistency
(e.g., removing circular segmentation)



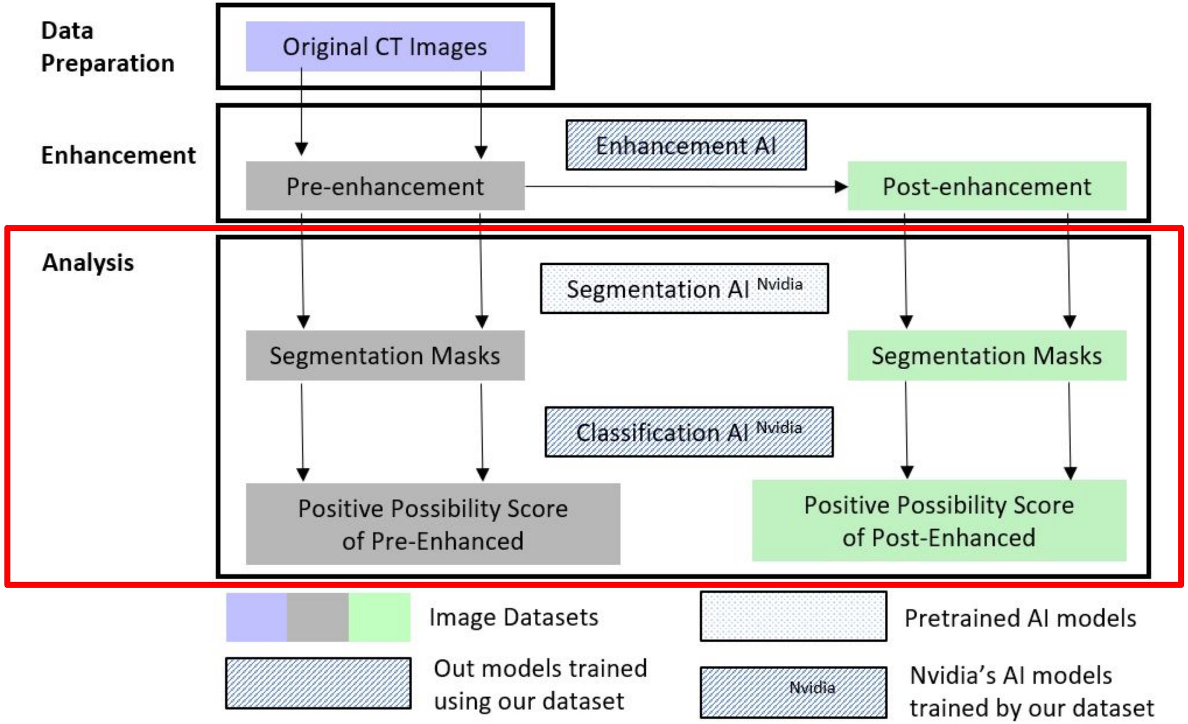
Workflow for *ComputeCOVID19+*



Enhancement (DenseNet and Deconvolution Network, a.k.a DDnet)



Workflow for *ComputeCOVID19+*



Analysis

Segmentation AI

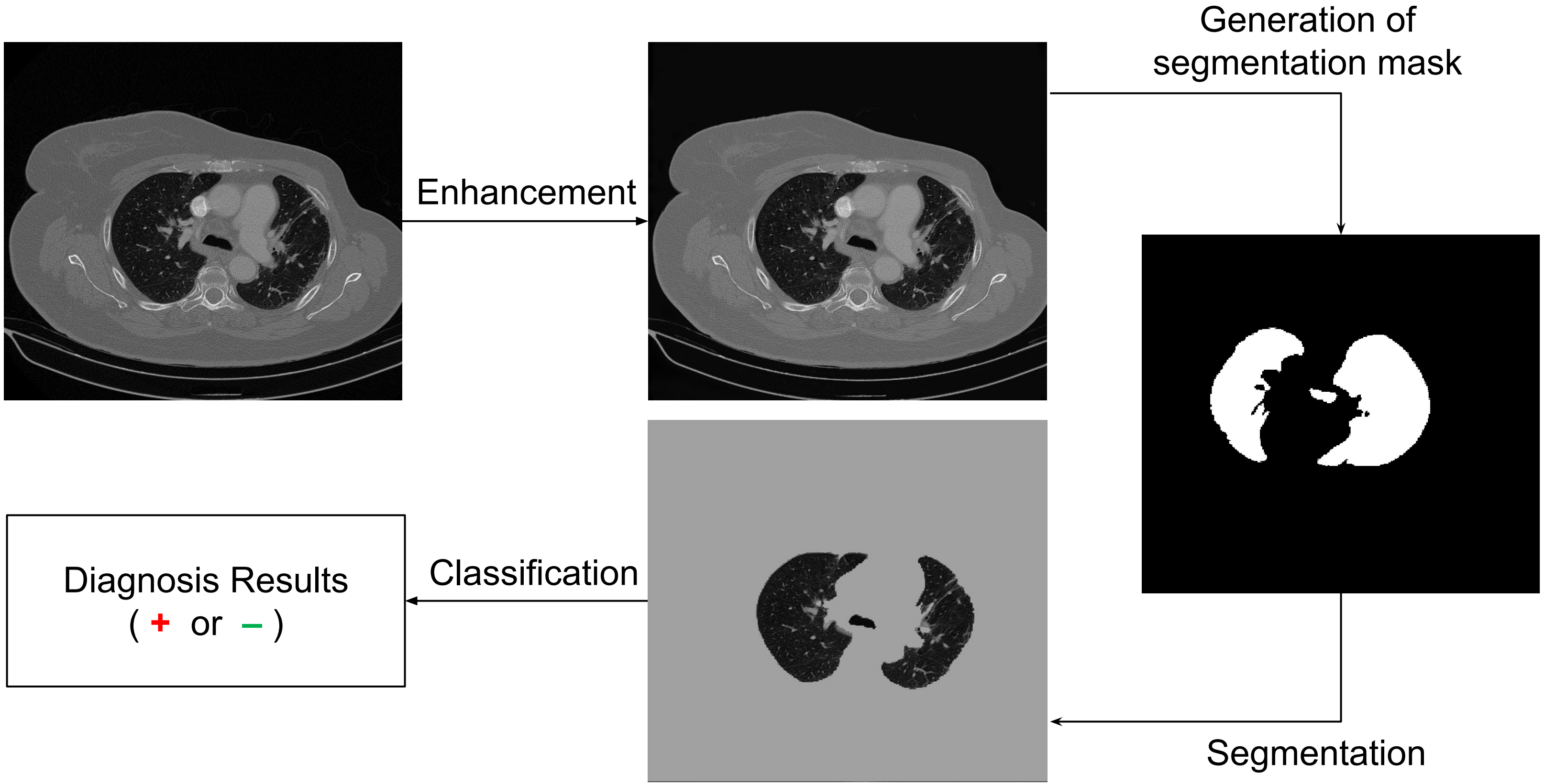
- Generates segmented images for improved interpretation of chest CT scans.
- Uses a 3D anisotropic hybrid network (i.e., 3D AH-Net).

Classification AI

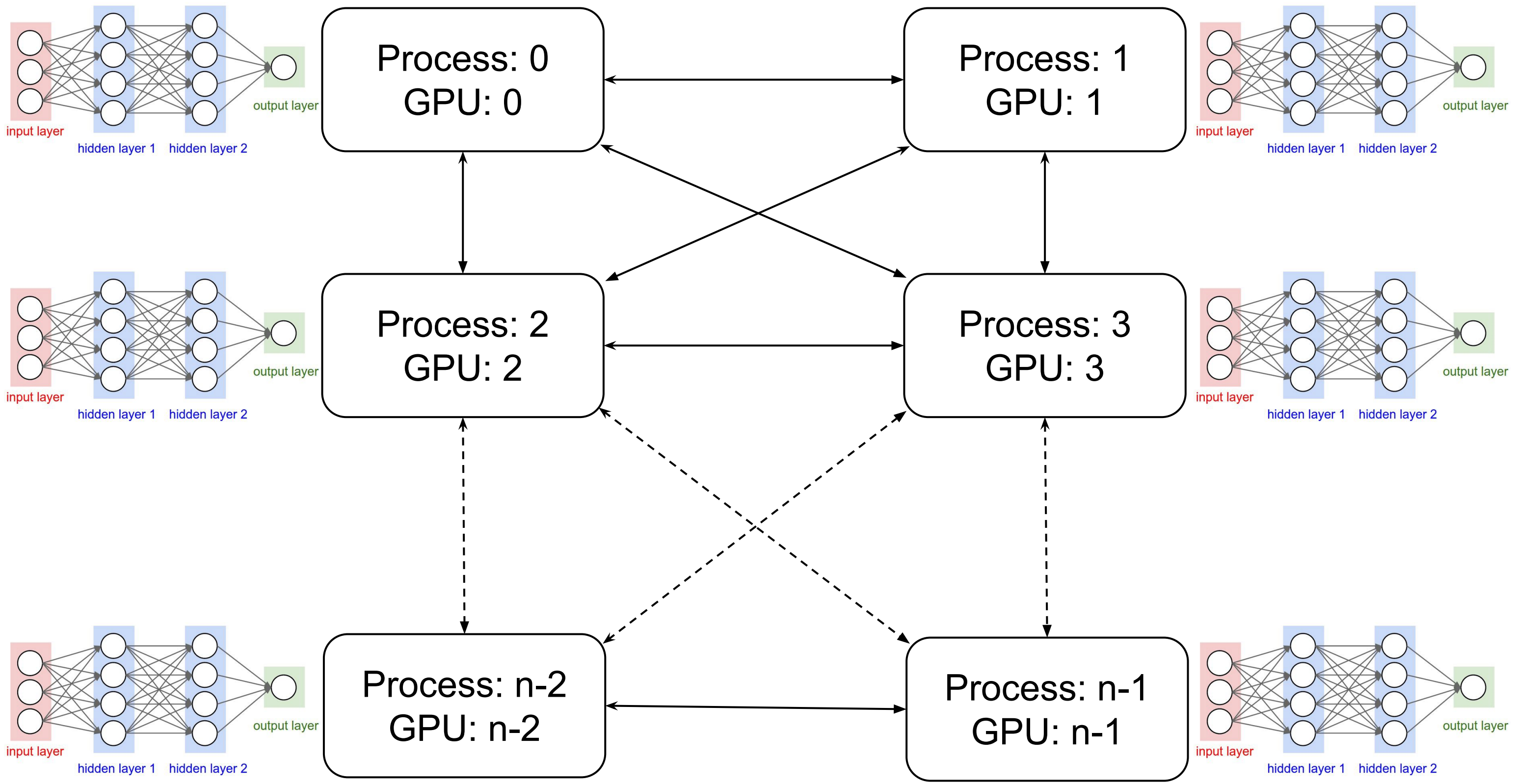
- Classifies the CT scan as + or – (based on features present).
- Uses a 3D DenseNet-121 network.

S. Harmon et al. 2020. Artificial Intelligence for the Detection of COVID-19 Pneumonia on Chest CT using Multinational Datasets. Nature Communications 11, 1 (Dec. 2020), 1–7.

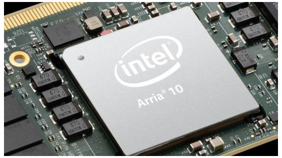
ComputeCOVID19+ Workflow



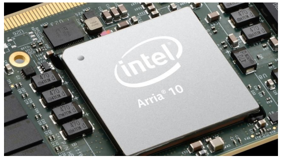
Optimizing Parallel Training



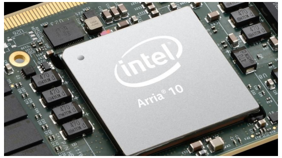
Inference on Heterogeneous Platforms

Platform	 **CPU**	 **GPU**	 **FPGA**
Strengths	• Multitasking with programming ease	• Massively parallel processing capability • Energy efficiency	• Configurable hardware for specific applications • Low power consumption • Energy efficiency
Weaknesses	• Supports limited parallelism	• High power consumption • Poor performance with irregular applications	• Relatively difficult to program • Operates at low frequency

Inference on Heterogeneous Platforms

Platform	 **CPU**	 **GPU**	 **FPGA**
Strengths	• Multitasking with programming ease	• Massively parallel processing capability • Energy efficiency	• Configurable hardware for specific applications • Low power consumption • Energy efficiency
Weaknesses	• Supports limited parallelism	• High power consumption • Poor performance with irregular applications	• Relatively difficult to program • Operates at low frequency

Inference on Heterogeneous Platforms

Platform	 **CPU**	 **GPU**	 **FPGA**
Strengths	• Multitasking with programming ease	• Massively parallel processing capability • Energy efficiency	• Configurable hardware for specific applications • Low power consumption • Energy efficiency
Weaknesses	• Supports limited parallelism	• High power consumption • Poor performance with irregular applications	• Relatively difficult to program • Operates at low frequency

Optimizing Parallel Inference

Application-specific optimizations	Architecture-specific optimizations	FPGA-specific optimizations
• Kernel refactorization (Optimizing deconvolution by reducing memory bandwidth requirement)	• Memory prefetching • Loop unrolling • Vectorization (inter-thread)	• Compute-unit replication • Vectorization (intra-thread) • Dedicated kernels • Runtime reconfiguration

Training Performance

# Nodes (i.e. # GPUs)	Batch Size	# Epochs	Training Runtime (hh:mm:ss)	MS-SSIM (Avg %)
1	1	50	15:14:46	98.71
4	8	50	2:27:49	96.35
4	8	100	4:58:52	96.30
4	16	50	2:07:58	95.18
8	8	50	2:21:49	95.46
8	8	100	4:43:26	95.78
8	32	50	1:17:25	92.04
8	64	50	1:12:24	88.02

Decreasing accuracy with batch size

Training Performance

# Nodes (i.e. # GPUs)	Batch Size	# Epochs	Training Runtime (hh:mm:ss)	MS-SSIM (Avg %)	
1	1	50	15:14:46	98.71	
4	8	50	2:27:49	96.35	
4	8	100	4:58:52	96.30	
4	16	50	**2:07:58**	**95.18**	~7x speedup
8	8	50	2:21:49	95.46	
8	8	100	4:43:26	95.78	
8	32	50	1:17:25	92.04	
8	64	50	1:12:24	88.02	

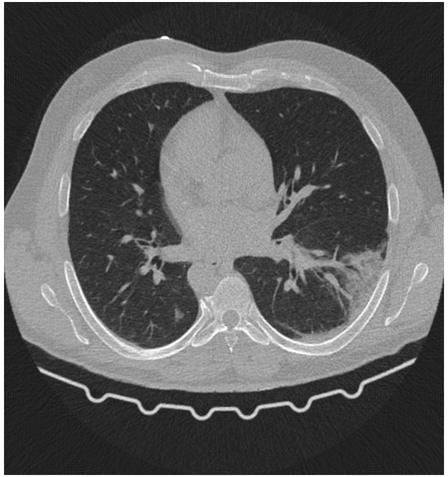
Inference Performance

Platform	Core Count	Maximum Bandwidth (GB/s)	Maximum Frequency (MHz)	PyTorch Runtime (seconds)	OpenCL Runtime (seconds)
Nvidia V100 GPU	5120 (CUDA cores)	900	1380	**0.22**	**0.10**
Nvidia P100 GPU	3584 (CUDA cores)	732	1328	0.73	0.25
AMD Radeon Vega Frontier GPU	4096 (Streaming proc.)	480	1600	–	0.25
Nvidia T4 GPU	5120 (CUDA cores)	320	1590	1.29	0.29
Intel Xeon Gold 6128 CPU	24 (CPU cores)	119	3400	5.52	1.64
Intel Arria 10 GX1150 FPGA	2 (Compute units)	< 3	184	–	16.74

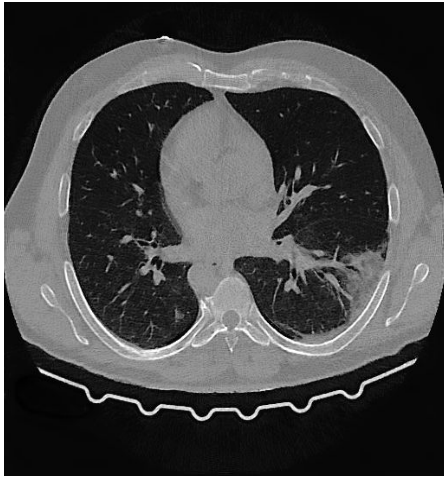
New runtime on FPGA: 0.5 seconds!

(–) : The PyTorch implementation is not portable to this platform.

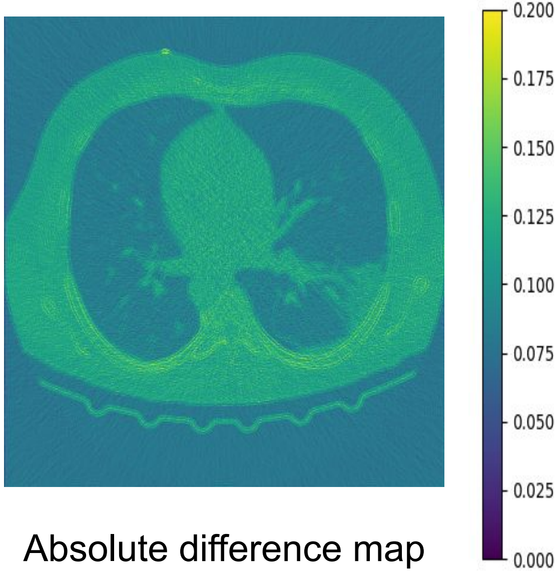
Evaluating Accuracy (Enhancement)



Quarter X-ray dose image



Enhanced image



Absolute difference map

Evaluating Accuracy (Enhancement)

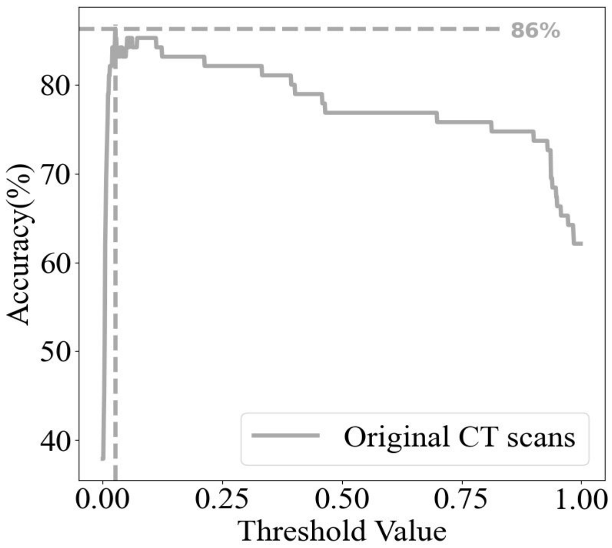
High-Quality Target Image vs.	**Mean Square Error** (lower is better)	**Multi-scale Structural Similarity Index (%)** (higher is better)
Low-Quality Input Image	0.00715	96.2
Enhanced Image	0.00091	98.7

$$MSE(x, y) = (\Sigma(x-y)^2)^{1/2}$$

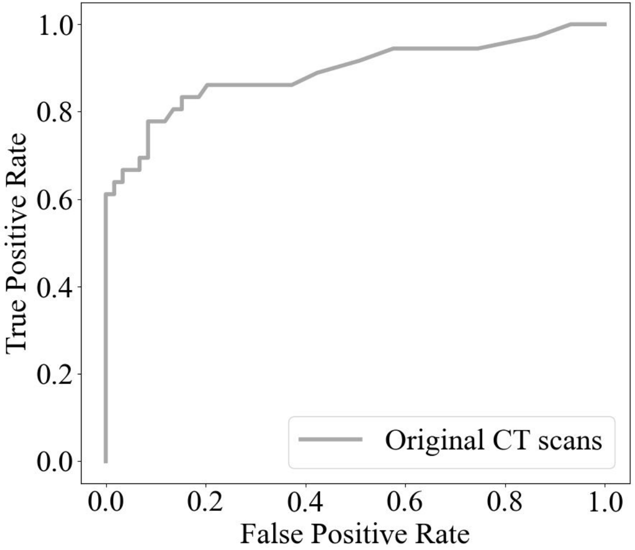
$$MS\text{-}SSIM(x, y) = f\,(luminance(x,y) * contrast(x, y) * structure(x,y))$$

Evaluating Accuracy (Analysis)

Accuracy



ROC Curve



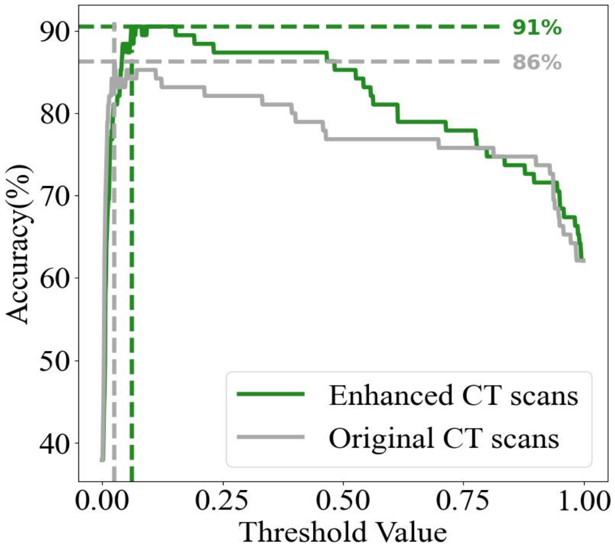
	Analysis
Accuracy	86%
ROC-AUC	0.890

INTERNATIONAL CONFERENCE ON PARALLEL PROCESSING

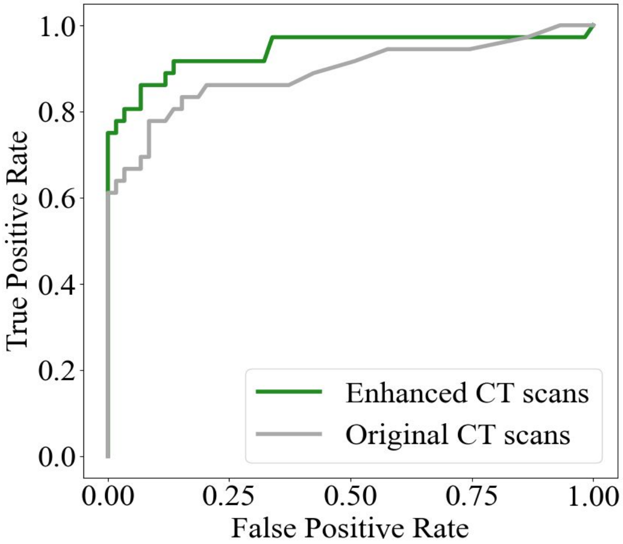
acm In-Cooperation sighpc

Evaluating Accuracy (Enhancement + Analysis)

Accuracy



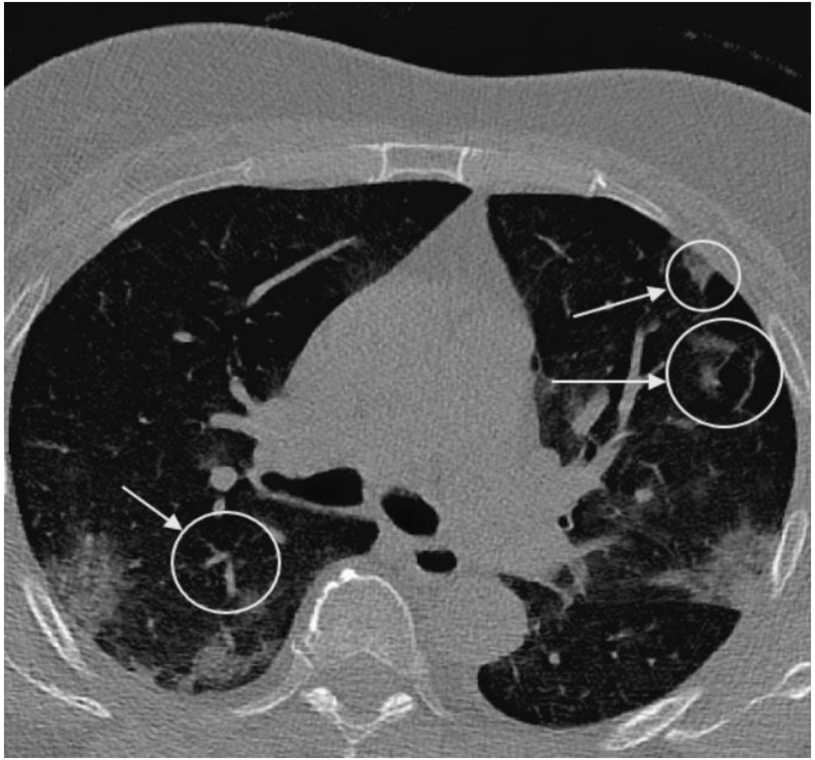
ROC Curve



	Analysis	Enhancement + Analysis	Improvement
Accuracy	86%	91%	5%
ROC-AUC	0.890	0.942	7%

Impact of Enhancement AI

Original Image



Enhanced Image

Conclusion

- ComputeCOVID19+ (https://github.com/vtsynergy/DL-FACT)
  - **Improvement in diagnosis accuracy** via enhanced CT imaging
  - **Faster turnaround time** due to better accessibility and due to less dependency on materials and labor.

Acknowledgements
1. NSF IIS-2027607
2. Cynthia McCollough, Mayo Clinic, and AAPM
3. VT Advanced Research Computing

Framework	CT Scans Pre-processing		2D/3D Classifier	Data Labelling	Hardware for Inference		
	Image Enhancement	Image Segmentation			CPU	GPU	FPGA
Compute-COVID19+	✔	✔	3D	Optional	✔	✔	✔
M-Inception	✕	✔	2D	Manual	–	–	✕
DRE-Net	✕	✔	2D	Manual	–	–	✕
DeCoVNet	✕	✔	3D	Optional	–	✔	✕