Advanced Foundations for Machine Learning Course Project

Physics Informed representation learning for COVID-19 Classification: An analysis of Feature separability and Latent space constraints

Aarya Upadhya, PES1UG23AM006
Anshull M Udyavar, PES1UG23AM057
Abhay H Bhargav, PES1UG23AM008
A Haveesh Kumar, PES1UG23AM001

Submission Checklist

PES UNIVERSITY ONLINE

No	Feature Description	Drive Shared (Y/N)
1	Code notebook	Y
2	Dataset or dataset source	Y
3	This PPT	Y
4	The 5 mins Video presenting your paper	Y
5	Brief Project Report in IEEE Format	Y

"The conflict : Interpretability vs Shortcut learning"

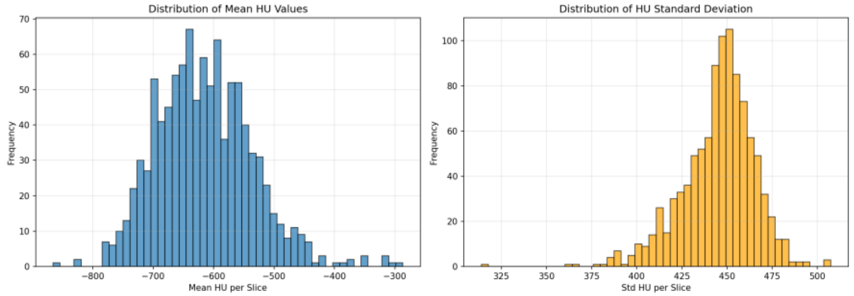
- Deep learning models often achieve high accuracy (90%+) but lack the clinical trust due to the "black box" nature
- Research shows that these models often learn shortcuts rather than pathology
- ***Our Hypothesis***: By constraining a model with 14 strict radiological physics attributes, we can force it to learn real medicine, and avoid shortcuts

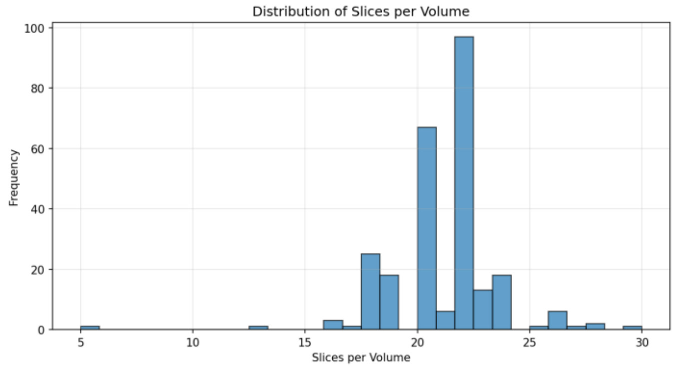
Uniqueness

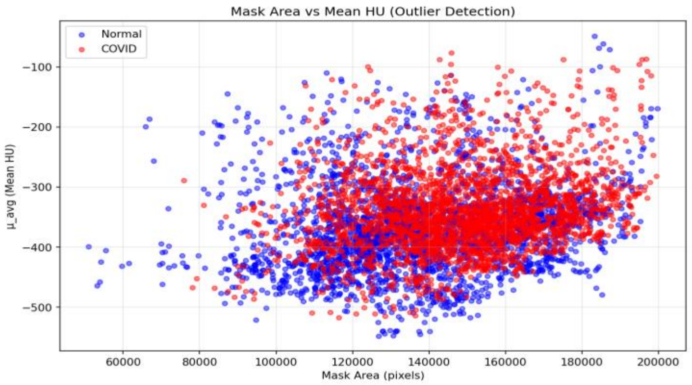
- ***Leak free splitting***: we split the data by Patient volume, not by slice ensuring no data leakage

- ***Physics Verification:*** We implement a pipeline to extract 13 different HU features(HU, gradient, texture)

- ***Sanity checks:*** We performed over 9 sanity checks, confirming that out data align with real world physics







Dataset

Dataset Size

- Used **5,000 CT slices** in total.
- Balanced: **2,500 COVID** and **2,500 Normal** lung CT images.
- Comes from a larger dataset containing scans from **over 1,000 patients**.

Dataset Attributes

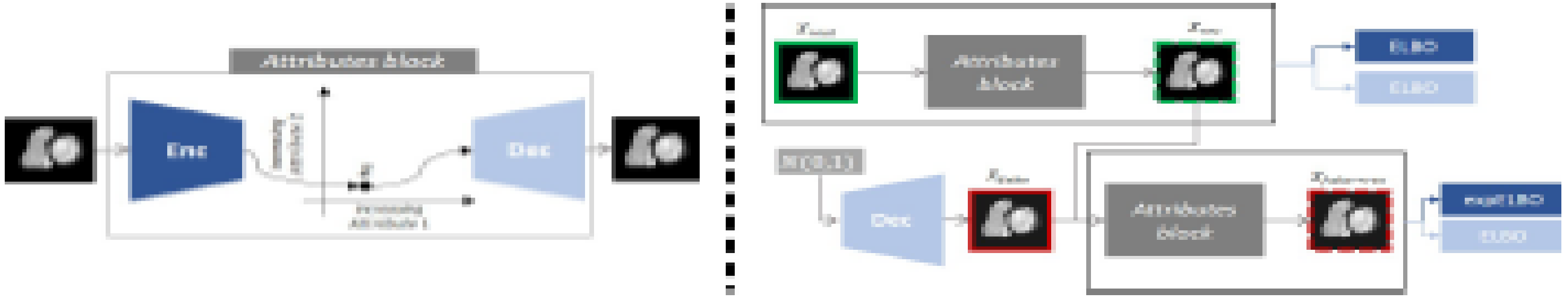
- Each sample is a **2D CT chest slice**.
- Includes variations in **scan quality, slice thickness, and acquisition settings**.
- Contains **pixel-level intensity values (Hounsfield Units)** that enable extraction of physics-based features.
- You generated **14 engineered attributes** per image (HU stats, texture, shape, gradient features) for physics-informed learning.

Dataset Source

- Sourced from **MosMedData**, a **public COVID-19 chest CT dataset** released by medical institutions in Moscow.
- Provided under an **open-access license** for research during the COVID-19 pandemic.

Overall Design or Approach

- # Approach :
- STAGE 1: data preprocessing and sanity checks
- STAGE 2: data exploration with drawn graphs
- STAGE 3: ARSIVAE with 1 physics
- STAGE 4: ARSIVAE with 14 radiological physics features
- # Design -



(a) Attributed-VAE (b) Proposed method

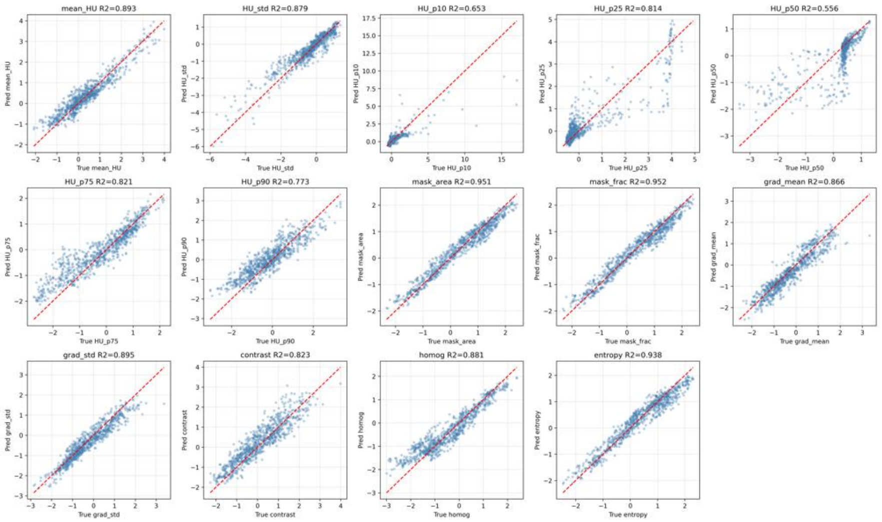
Fig. 1: (a) Principle of the attributed regularization proposed by Pati et al. [3]. The loss is composed of a classic VAE loss and an attribute regularization term described in the methods section. (b) Global network framework of the proposed method: Attri-SIVAE based on the Soft Introspective VAE [4]. Our contribution relies on integrating the attributes regularization term into the framework.

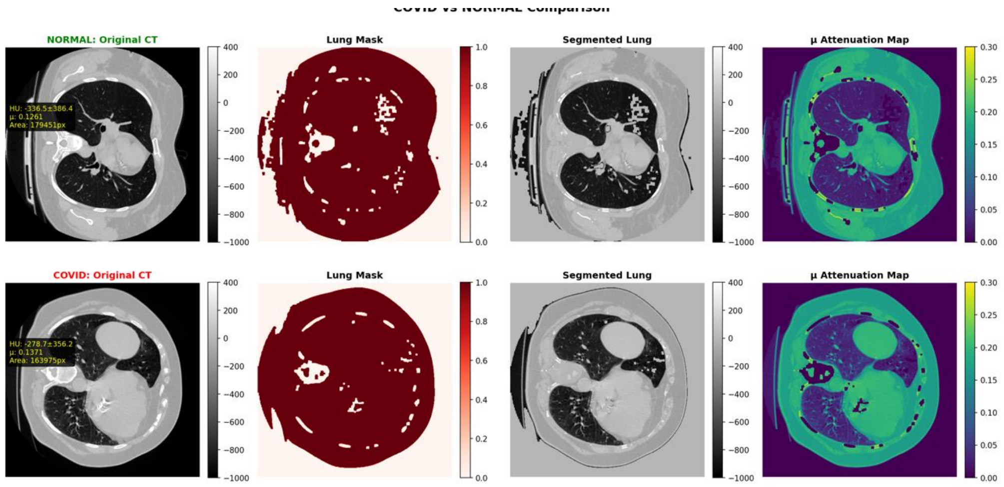
Model Success – Physics alignment

- ***Technical Success***: The ARSIVAE successfully learned the laws of radiology

- ***Metric***: We achieved an average R^2 value of 0.89 across all 14 physics attributes

- ***Implication***: The model successful compressed the images into a physically meaningful latent space. It understands density and texture via the HU units



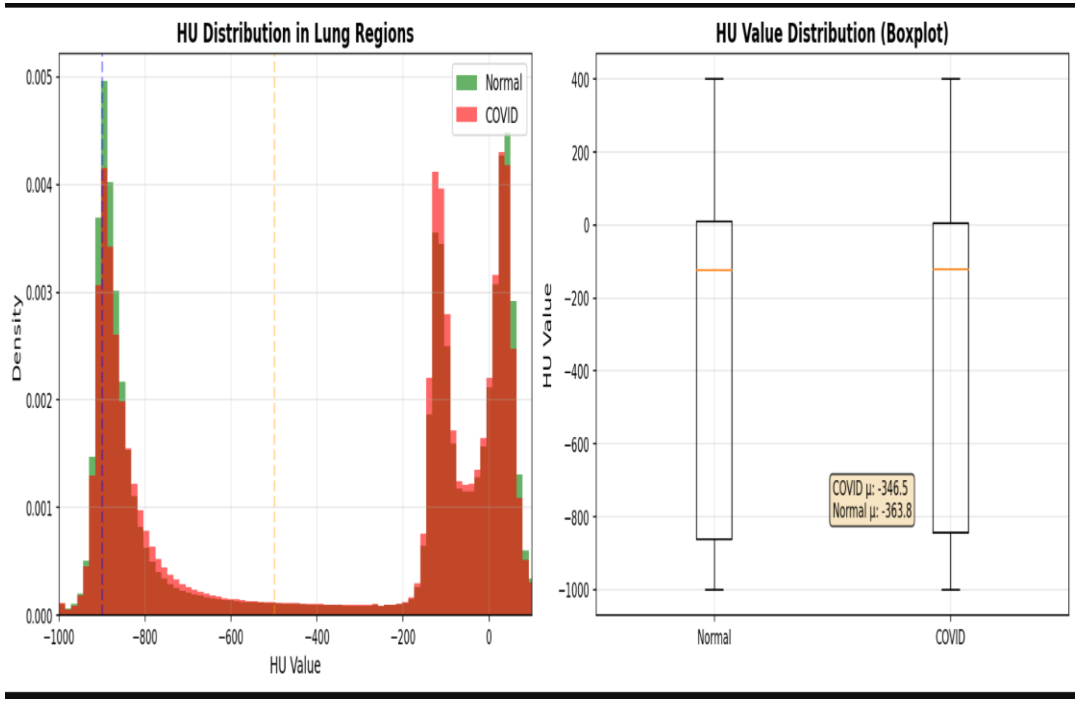


The discovery – “The spatial Information Gap”

- ***The ceiling***: the perfect Physics understandings , classification accuracy plateaued at 70%(on Gridsearch and linear probe).

- ***The Cause***: Statistical analysis reveals around 35% to 70% overlap in the global physics between COVID and Normal cases

- ***The Finding***: Global averages wash out localized pathologies like GGO. A mild COVID lung statistically resemble a healthy lung

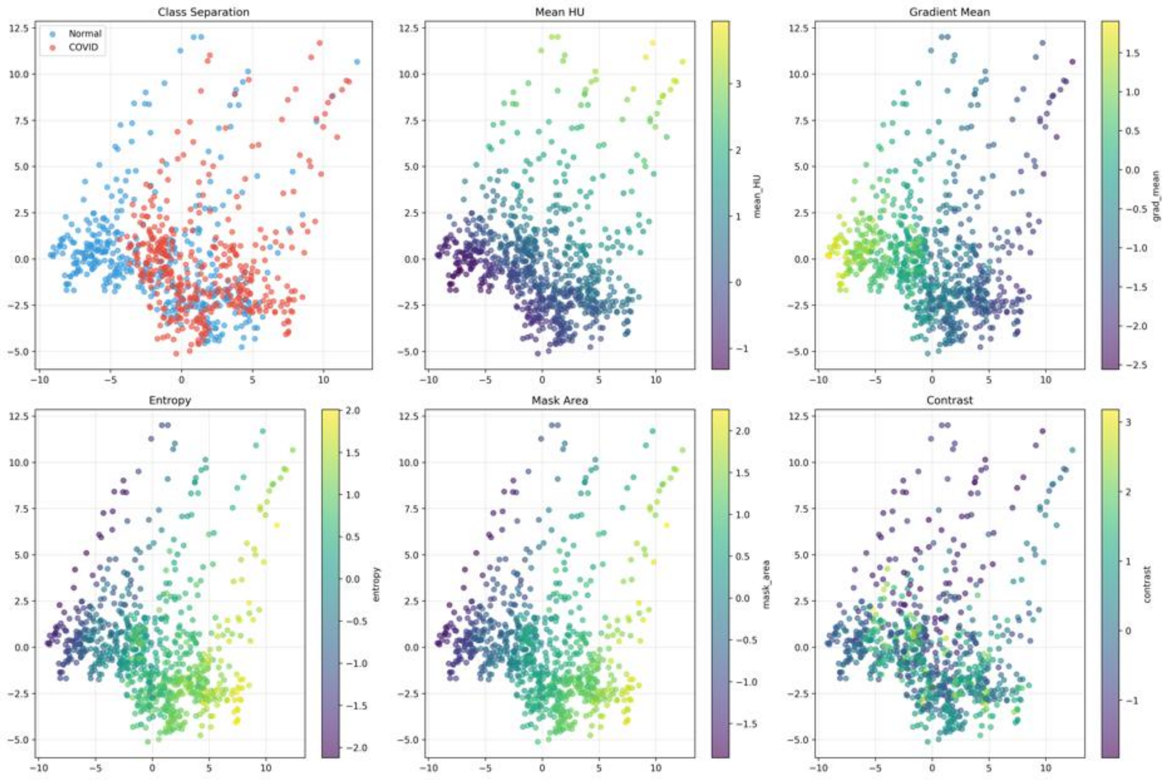


Final Results

- ***Visualizing the limit:*** The latent space shows intertwined clusters rather than clean separation

- ***Interpretation***: This proves that without spatial localization, the classes are not linearly separable

- ***Scientific Value:*** If we used a standard CNN, it might have separated these falsely using “shortcuts”. Our model's confusion is actually an honest representation of the data’s ambiguity



Features : Done vs. Remaining to be done

No	Description	Done or To be Done ?
1	Data extraction and preprocessing	Done
2	Physics feature extraction with HU units	Done
3	ARSIVAE with 1 physics attribute	Done
4	ARSIVAE with 14 physics features and downstream classification	Done
5	Hybrid model Spatial extraction + Physics extraction model	To be Done

Features : Who did what

No	Feature Description	Contributed By
1	Model building and interpretation	Aarya Upadhya
2	Model building and interpretation	Anshull M Udyavar
3	Sanity checks and data analysis	Abhay H Bhargav
4	Preprocessing and data loader	A Haveesh Kumar

Quantity and Quality of Work

No	Code Functionality	% Complete	Runs w/o Issues (Y /N)	State minor issues
1	Data extraction and preprocessing	100%	Y	No minor issues . Patient leakage resolved
2	Physics Feature engineering	100%	Y	Clipping and masking lungs to get right HU units resolved
3	ARSIVAE 1 physics attribute (test)	100%	Y	Global averaging
4	ARSIVAE 14 physics	100%	Y	Separability issue
5	Classification on physics latent space	100%	Y	Low performance

Top Few Learnings from this Project

No	Description
1	**Physics-based global CT features are informative but not fully discriminative**, showing 35–40% inherent class overlap.
2	**Interpretability and accuracy can conflict** — the ARSIVAE achieved high physics alignment (R^2 = 0.89) but only moderate classification performance.
3	**Multi-objective training matters** — staged annealing prevented latent space collapse and balanced physics + classification goals.
4	**Spatial information is essential** — global HU/texture statistics miss localized COVID patterns like GGOs.
5	**Hybrid models are the future** — combining physics priors with spatial deep learning can bridge interpretability and performance.

Top Unresolved Challenges

No	Description
1	Down stream classification via the physics latent space
2	Downstream separation of the classes into different clusters
3	Understanding the need for spatial features to resolve separation Gap early

References, if any

No	Paper title	Year of publication
1	R. R. Jain, A. S. Kanhere, C. T. M. H. L. L. Cabral, and G. Hamarneh, ”ARSIVAE: Attribute-Regularized Soft Introspective Variational Autoencoder for Medically Explainable Classification of Chest CT Scans,” arXiv preprint arXiv:2406.08282, 2024.	2024
2	P. Perdikaris et al., ”Physics-informed deep learning for surrogate modeling of urban thermal dynamics,” arXiv preprint arXiv:2312.08915, 2023.	2023
3	D. P. Kingma and M. Welling, ”Auto-encoding variational bayes,” in Proc. Int. Conf. Learn. Represent. (ICLR)	2014
4	I. Higgins et al., ”beta-VAE: Learning basic visual concepts with a constrained variational framework,” in Proc. Int. Conf. Learn. Represent. (ICLR)	2017

References, if any

No	Paper title	Year of publication
5	L. Li et al., ”Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT,” Radiology, vol. 296, no. 2, pp. E65-E71	2020
6	A. J. DeGrave, J. D. Janizek, and S. Lee, ”AI for radiographic COVID19 detection selects shortcuts and fails to generalize,” Nature Machine Intelligence, vol. 3, no. 7, pp. 610-619	2021
7	J. R. Zech et al., ”Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study,” PLoS Medicine, vol. 15, no. 11, p. e1002683	2018
8	A. Holzinger et al., ”Causability and explainability of artificial intelligence in medicine,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 9, no. 4, p. e1312	2019

PES UNIVERSITY ONLINE

References, if any

No	Paper title	Year of publication
9	R. R. Selvaraju et al., ”Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 618-626.	2017

THANK YOU

November 2025

Department of Computer Science and Engineering in AI &ML