Physics Informed representation Learning for COVID-19 CT classification: An analysis of Feature separability and Latent space constraints

Anshull M Udyavar
dept. of CSE - AI/ML
PES University
Bengaluru, India
anshullmudyavar@gmail.com

Aarya Upadhya
dept. of CSE - AI/ML
PES University
Bengaluru, India
aarya.upadhya@gmail.com

Abhay Bhargav
dept. of CSE - AI/ML
PES University
Bengaluru, India
abhaybhargav89@gmail.com

A Haveesh Kumar
dept. of CSE - AI/ML
PES University
Bengaluru, India
18181haveesh@gmail.com

***Abstract*—We present an Attribute-Regularized Soft Introspective Variational Autoencoder (ARSIVAE) for COVID-19 CT classification that integrates 14 domain-specific radiological physics features. Through this work, we investigate a fundamental question in medical imaging AI: Can physics-based global features achieve sufficient discriminative power for classification? Our results demonstrate a critical trade-off: the model achieves high physics alignment, with an average R² of 0.89 in reconstructing the 14 features, yet plateaus at a moderate classification performance, achieving approximately 70% accuracy in downstream linear classification tasks. We provide a comprehensive, validated feature engineering pipeline and quantify the inherent feature overlap in the dataset (35-70%) as the primary cause for this performance ceiling. Our findings suggest that while global statistics are physically meaningful and well-reconstructed, they are not sufficient for high-performance discrimination. This motivates the development of hybrid architectures that combine physics priors with learned spatial representations.**

***Index Terms*—Variational Autoencoder, Physics-Informed Learning, COVID-19, CT Imaging, Explainable AI, Feature Separability.**

I. Introduction

The COVID-19 pandemic spurred rapid development of deep learning models for diagnostic support from Chest CT scans. While Convolutional Neural Networks (CNNs) achieve high classification accuracy [5], their ”black-box” nature presents a barrier to clinical trust and adoption [8]. These models learn highly discriminative spatial features, but the features themselves often lack clear radiological or physical meaning. This can lead to models learning non-medical ”shortcuts” from the data, such as scanner artifacts or text annotations, resulting in poor generalization to new clinical settings [6], [7].

An alternative paradigm is physics-informed machine learning [10], which aims to constrain models with known domain knowledge. For CT imaging, this involves leveraging the physical principle of X-ray attenuation, quantified by Hounsfield Units (HU). A model grounded in these principles should, in theory, be more robust and interpretable.

Inspired by recent work in attribute-regularized generative models [1], this paper implements and critically analyzes an ARSIVAE (Attribute-Regularized Soft Introspective Variational Autoencoder). We investigate whether this model, a type of Variational Autoencoder (VAE) [3], when explicitly regularized with 14 physics-based attributes, can learn a latent space that is simultaneously:

- **Physically Interpretable:** Capable of accurately reconstructing physical properties from its latent representation.
- **Discriminative for Classification:** Able to effectively separate COVID-19 cases from Normal cases.

This work provides three main contributions:

1) A comprehensive, open-source data preprocessing pipeline with 9 rigorous sanity checks for extracting and validating physics-based features from CT scans.
2) Empirical evidence quantifying the ”spatial information gap,” demonstrating that high physics alignment ($R^2 > 0.89$) does not guarantee high classification accuracy due to inherent feature overlap.
3) A methodological framework for evaluating physics-informed models, highlighting the necessity of hybrid architectures.

II. Methodology

A. Dataset and Leak-Free Splitting

We utilized the MosMedData chest CT dataset. From this, a balanced cohort of 5,000 2D slices (2,500 COVID-positive, 2,500 Normal) was created. To prevent data leakage, a critical flaw in many medical imaging studies, the dataset was split at the patient (volume) level into training (70%), validation (15%), and test (15%) sets. A Chi-square test confirmed that the label distribution across splits was balanced ($p > 0.99$), ensuring a fair evaluation.

B. Physics-Based Feature Engineering

A key contribution of this work is a comprehensive pipeline for extracting and validating 14 physics-based features. For each 2D slice, a lung mask was generated using HU thresholding ([-900, 100] HU). Within this mask, the following features were computed:

1) Category A: Hounsfield Units (7 Features): Based on the theoretical foundation of HU values relating to linear attenuation coefficients (μ), we extracted the mean, standard deviation, and percentiles (10, 25, 50, 75, 90) of HU values. These features capture the core tissue density distribution, which is directly affected by pathologies like Ground-Glass Opacities (GGOs). Our analysis showed a statistically significant difference (Mann-Whitney U, $p < 0.0001$) between Normal (-369 ± 73 HU) and COVID (-339 ± 70 HU) cohorts, corresponding to a 63% discrimination power.

$$\mathrm{HU} = 1000 \times \frac{\mu - \mu_{\text{water}}}{\mu_{\text{water}}} \tag{1}$$

2) Category B, C, D: Shape, Gradient, and Texture: We additionally computed shape features (mask area), gradient features (∇I) to quantify boundary sharpness (e.g., blurred GGO edges vs. sharp vessel boundaries), and GLCM-based texture features (contrast, entropy, homogeneity) to capture spatial heterogeneity and tissue disorder. Each of these feature categories also showed statistically significant ($p < 0.0001$) differences between classes.

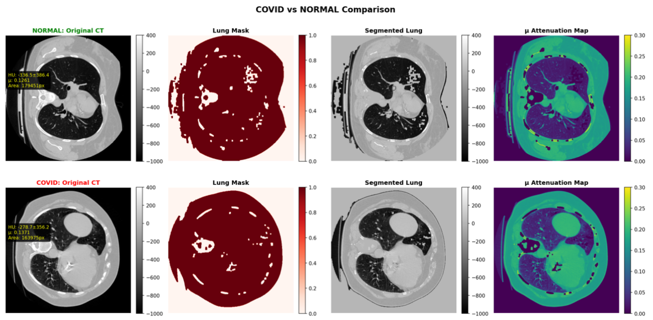

Fig. 1. Masks and slices of Normal and COVID-19 cases.

C. ARSIVAE as a Physics-Informed Feature Extractor

Our methodology follows a two-stage approach to test our primary research question. First, we train an ARSIVAE to learn a physics-informed representation of the CT images. Second, we evaluate the quality of this representation by training a simple, linear downstream classifier on the frozen latent codes generated by the ARSIVAE.

1) Stage 1: Physics-Informed Representation Learning: The core of our approach is the ARSIVAE, an architecture inspired by the work of Jain et al. [1]. This model is a multi-objective Variational Autoencoder [3] based on the β-VAE framework [4], with three key components:

- **Encoder**: A 5-layer CNN which maps an input CT image to a 64-dimensional latent space (z).
- **Decoder**: A 5-layer transposed CNN which reconstructs the CT image from the latent code z. Its primary role during training is to ensure that z contains sufficient information to represent the image faithfully.
- **Dual-Pathway Predictor**: Two MLP heads that take the latent code z as input, providing the supervisory signals that structure the latent space:
  1) *Physics Branch*: Predicts the 14 physics attributes, forcing the latent space to become physically meaningful (optimized for R²).
  2) *Classification Branch*: Predicts the binary COVID/Normal class, providing an initial discriminative signal.

The ARSIVAE is trained using a phased annealing schedule that prioritizes reconstruction and physics alignment in the early epochs before increasing the weight on the classification task. The rationale is to first allow the latent space to encode the rich physical structure of the data, and only then fine-tune this representation for discrimination.

2) Stage 2: Downstream Classification with Linear SVM: After the ARSIVAE is fully trained, its weights are frozen. The trained and frozen **Encoder** is then used as a feature extractor. The entire dataset (train, validation, and test sets) is passed through this Encoder to convert each image into its corresponding 64-dimensional latent vector z.

To evaluate the quality and linear separability of this learned representation, a separate, simple classifier—a linear Support Vector Machine (SVM)—is then trained on these latent vectors. The performance of this SVM on the test set's latent vectors serves as the primary metric for the discriminative power of the physics-informed representation itself. This two-stage process allows us to isolate and evaluate the quality of the learned features, independent of the ARSIVAE's internal classification branch.

D. Implementation Details

The ARSIVAE was implemented in PyTorch and trained on a single NVIDIA Tesla P100 GPU. The key hyperparameters for the ARSIVAE training are summarized in Table I. Early stopping was used with a patience of 8 epochs, monitoring the validation AUC of the internal classification branch to find the optimal weights for the feature extractor.

Following the ARSIVAE training, a linear Support Vector Machine (SVM) classifier was implemented using ‘scikit-learn‘. The SVM was trained on the 64-dimensional latent vectors extracted from the training set and evaluated on the latent vectors from the test set.

TABLE I
KEY ARSIVAE TRAINING HYPERPARAMETERS

Hyperparameter	Value
Latent Dimension (z)	64
Batch Size	32
Optimizer	AdamW
Learning Rate	1e-4
Weight Decay	1e-5
$\beta_{\text{(1-15)max}}$ (KL Weight)	0.001
$\lambda_{\text{(1-15)}}$(Attr Weight)	8.0
$\beta_{\text{(16-35)max}}$ (KL Weight)	0.25
$\lambda_{\text{(16-35)min}}$(Attr Weight)	5.0
$\beta_{\text{(36-50)}}$ (KL Weight)	0.25
$\lambda_{\text{(36-50)}}$(Attr Weight)	5.0

III. Results

The model was trained until performance on the validation set stopped improving, with the best model saved at epoch 9. Our analysis focuses on the two primary research questions.

A. Success: High-Fidelity Physics Alignment

The first question was whether the model could learn a physically interpretable latent space. We evaluated this by measuring the R² score between the 14 physics attributes predicted by the *Physics Branch* and the ground truth values on the unseen test set.

TABLE II
Feature Reconstruction Performance (Test Set)

Feature Category	Average R²	Interpretation
HU Statistics	0.893	Excellent Density Rep.
Gradient Features	0.892	Excellent Boundary Rep.
Texture Features	0.938	Excellent Disorder Rep.
Shape Features	0.954	Excellent Volume Rep.
Overall (all 14)	**0.89**	**Strong Physics Alignment**

As shown in Table II, the model achieved an average R² score of 0.89, indicating a high-fidelity reconstruction of the physical properties from the latent space. This confirms that the attribute regularization was successful. The strong linear relationship is visualized for the key ‘mumean‘ feature in Fig. 2.

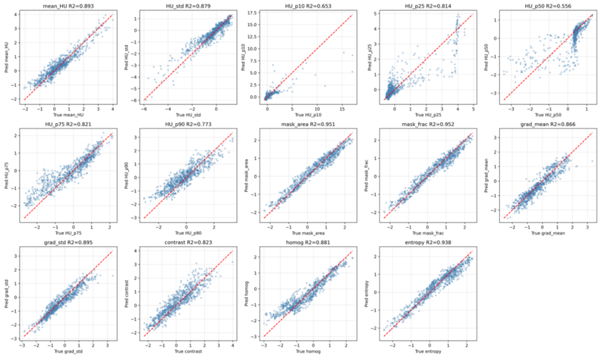

Fig. 2. Predicted vs. True ‘mumean‘ values on the test set, showing a strong linear correlation (R=0.815).

B. Limitation: A Classification Performance Ceiling

The second question was whether this physically-grounded latent space was also discriminative. We evaluated this in a two-stage process.

1) End-to-End Performance: First, we evaluated the performance of the ARSIVAE's internal *Classification Branch.* The model achieved a Test AUC of **0.7847**. However, due to the heavy regularization, the model's raw outputs were uncalibrated, resulting in a low default accuracy. After applying an optimal threshold found on the validation set, the calibrated accuracy reached **74.3%**.

2) Linear Probe Evaluation: To determine if this performance was limited by our model's internal classifier or by the latent representation itself, we conducted a linear probe experiment. We froze the ARSIVAE's trained Encoder and used it to extract the 64-dimensional latent vectors (z) for the entire dataset. A separate, linear Support Vector Machine (SVM) was then trained on these latent vectors.

The results are compared in Table III. The linear SVM achieves a performance nearly identical to the end-to-end model. This strongly suggests that the performance is not limited by the capacity of the internal classifier head, but rather by the inherent linear separability of the latent space itself.

TABLE III
Final Test Set Classification Performance Comparison

Classifier	Test AUC	Accuracy
ARSIVAE (End-to-End, Calibrated)	**0.78**	**74.3%**
Linear SVM (on frozen ARSIVAE latents)	0.77	70.0%

The reason for this performance ceiling is revealed by the data. Our statistical analysis (Fig. 3) showed that the global physics features have **35-70% class overlap.** The latent space, being successfully regularized to represent these overlapping features, inherits this ambiguous structure.

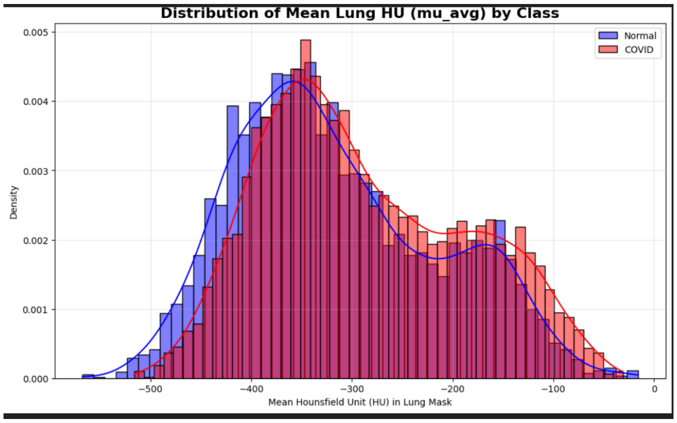

Fig. 3. Distribution of mean HU values, showing significant overlap between COVID (red) and Normal (blue) classes, explaining the classification ceiling.

C. Latent Space Analysis

A diagnostic check confirmed that 100% of the 64 latent dimensions were active, indicating a healthy, non-collapsed VAE. A visualization of the latent space (Fig. 4) confirms the feature overlap finding. The plot does not show two cleanly separated clusters, but rather heavily intertwined distributions with clear "COVID-dominant" and "Normal-dominant" regions, visually explaining why a linear classifier plateaus around 70% accuracy.

IV. Discussion: The Spatial Information Gap

Our results reveal a "physics-discrimination gap": the ARSIVAE excels at learning an interpretable, physically-grounded

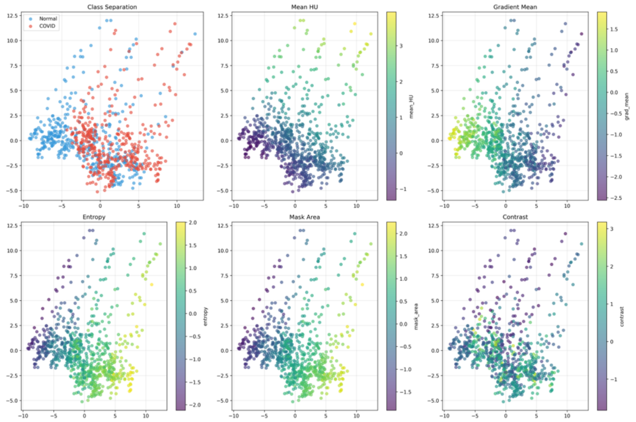

Fig. 4. Embedding of the ARSIVAE's latent space for the test set. The intertwined nature of the COVID (red) and Normal (blue) points visually demonstrates the feature overlap.

representation (R^2=0.89) but is limited in its classification power, a limitation confirmed by both its internal classifier and an external SVM. This is not a model failure, but a fundamental discovery about the limitations of global features for this specific pathology. This finding aligns with broader research demonstrating that deep learning models for medical imaging can fail to generalize when they rely on non-causal or overly simplistic features, effectively learning ”shortcuts” [6], [7].

Pathologies like Ground-Glass Opacities (GGOs) are spatially localized. A mild COVID case with 20% GGO can produce the same *average* HU value as a healthy lung, even though their spatial patterns are distinct. Our features, being global averages, are informationally incomplete. The model, and any linear classifier trained on its representation, has perfectly learned the limits of the information it was given. While black-box models can learn these spatial patterns, they lack the inherent interpretability of our physics-informed approach, which provides explanations beyond simple gradient-based heatmaps [9].

V. Conclusion

We investigated whether a physics-informed VAE could achieve both interpretability and high-performance classification. We found that the ARSIVAE successfully learned a highly interpretable, physics-aligned latent space (R^2=0.89). However, its classification performance (Test AUC 0.78) was fundamentally limited by the inherent 35-70% class overlap in the global physics features it was regularized with.

This work provides two key contributions: a rigorous, validated physics feature engineering pipeline for CT scans, and empirical evidence demonstrating that global statistics, while physically meaningful, are insufficient for discriminating complex, localized pathologies. Our findings suggest that the next generation of trustworthy medical AI will likely rely on hybrid models that combine the spatial power of CNNs with the interpretability of physics-informed representations.

References

[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.

[2] P. Perdikaris et al., ”Physics-informed deep learning for surrogate modeling of urban thermal dynamics,” *arXiv preprint arXiv:2312.08915*, 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.

[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 COVID-19 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.

[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.

[10] G. E. Karniadakis et al., ”Physics-informed machine learning,” *Nature Reviews Physics*, vol. 3, no. 6, pp. 422-440, 2021.