# VAE-Based Feature Extraction for Lung Nodule Diagnosis

Zhihan Liu
Universidad Autonoma de Barcelona
School of Engineering
Barcelona, Bellaterra, Spain
Zhihan.Liu@autonoma.cat

Debora Gil
Universidad Autonoma de Barcelona
Computer Vision Center
Barcelona, Bellaterra, Spain
debora@cvc.uab.es

Guillermo Torres
Universidad Autonoma de Barcelona
Computer Vision Center
Barcelona, Bellaterra, Spain
gtorres@cvc.uab.es

## Abstract

The manner with which CT scans are acquired, reconstructed, displayed and interpreted impacts radiologists' performance. Fundamental to the diagnosis of lung cancer in CT scans is the detection and interpretation of lung nodules. Motivated by these challenges, we propose an end-to-end framework that combines a Variational Autoencoder (VAE) and a Multi-Layer Perceptron (MLP) for the reconstruction and diagnosis of lung nodules in CT imaging. The framework is composed of three stages: (1) Preprocessing: preprocessing of CT images into VAE-compatible inputs; (2) VAE training: unsupervised training of VAEs and finding optimal hyperparameters; and (3) Joint training: joint fine-tuning of the VAE and MLP in a unified model. During training, we optimize VAE reconstruction using SSIM, MSE, and validation loss. An optimal hyperparameter group is selected based on a custom scoring function. Finally, in the joint training stage, we attach an MLP classifier to the trained encoder and fine-tune the entire model using a combination of reconstruction and classification losses. Our framework demonstrates strong diagnostic performance. These results suggest that the proposed VAE-MLP framework shows strong diagnostic performance. It effectively captures both structure and diagnostic features, offering a great solution for automated lung nodule analysis.

## CCS Concepts

• **Computing methodologies → Computer vision**.

## Keywords

Variational Autoencoder, Multi-Layer Perceptron, CT imaging reconstruction, Lung nodule detection

**ACM Reference Format:**
Zhihan Liu, Debora Gil, and Guillermo Torres. 2025. VAE-Based Feature Extraction for Lung Nodule Diagnosis. In *Proceedings of the International Workshop on Automotive and Medical Multimedia: Bridging the Gap Between Mobility and Healthcare (AMM '25), October 27–28, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3746271.3776606

## 1 Introduction

Lung cancer is the most commonly diagnosed cancer and the leading cause of cancer-related deaths globally [4]. In 2022 almost 2.5 million people were diagnosed with lung cancer and more than 1.8 million people died from the disease [8]. The chance of an effective treatment outcome and the likelihood of patient survival can be greatly improved with the early identification of lung cancer [6]. Therefore, accurate diagnosis of pulmonary nodules in CT imaging is a key task in clinical practice. Manually analyzing CT scans is time-consuming and labor-intensive and may have inconsistent diagnosis between doctors due to differences in experience or missing some features. Therefore, there is more need for automatic and exact tools to help doctors check lung nodules fast and right is growing.

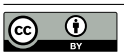
Deep learning algorithms, powered by advances in computation and very large datasets, have recently been shown to exceed human performance in visual tasks [1]. Although researchers have successfully applied CNNs for many medical image understanding applications like detection of tumors and their classification into benign and malignant [5], etc. Their reliance on labeled data, and lack of generative or reconstructive capabilities. This motivates the exploration of generative models such as Variational Autoencoders (VAEs), which can learn meaningful latent representations from unlabeled CT data, meanwhile providing image reconstructions that aid clinical insight.

Variational Autoencoders (VAEs) marry graphical models and deep learning [2]. As an unsupervised generative model, the VAE provides a principled way for performing variational inference utilizing an Autoencoder (AE) architecture [7]. This structure makes sure the model can learn continuous and interpretable latent representations of high-dimensional input images such as lung CT images. A related advantage of this structure is VAE can ability to control the distribution of the latent representation vector z, which can combine VAEs with representation learning to further improve the downstream tasks [7], such as classification, anomaly detection, and disease staging.

To further enhance the diagnosis function on lung nodule, we integrate the generative representations learned by the VAE with a model optimized focus on classification. Accordingly, we propose a hybrid framework that combines the representational ability of a VAE with the discriminative ability of a Multi-Layer Perceptron (MLP) classifier. This VAE-MLP framework combines the respective advantages of the two models: the VAE captures many interpretable latent features through unsupervised learning and image reconstruction, while the MLP classification optimizes for diagnostic accuracy. By jointly training the combined model end-to-end, we enable mutual refinement between the generative and discriminative models, leading to getting better reconstruction results and diagnostic results without sacrificing interpretability.

## 2 Methodology

Our proposed framework is composed of three important components: (1) CT image preprocessing, (2) unsupervised training of a



This work is licensed under a Creative Commons Attribution 4.0 International License.
*AMM '25, Dublin, Ireland*
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2053-6/2025/10
https://doi.org/10.1145/3746271.3776606

Variational Autoencoder (VAE), and (3) end-to-end joint training of the VAE encoder and an MLP classifier. The overall objective is to learn a latent representation that captures both structural patterns for image reconstruction and discriminative features for diagnosis of lung nodules. Both the VAE and MLP models are trained exclusively on the training set, while the validation set is used solely for performance evaluation without influencing model training or parameter adjustment.

## 2.1 Dataset

The experiments in this study are based on the LIDC-IDRI (Lung Image Database Consortium and Image Database Resource Initiative) dataset, a publicly available large-scale dataset for lung nodule analysis. This dataset was collaboratively developed by the National Cancer Institute (NCI) and is widely used in medical imaging research for lung cancer detection and diagnosis.

Each CT scan in LIDC-IDRI is accompanied by detailed nodule annotations from four experienced radiologists, who independently marked and rated nodules based on characteristics such as size, margin, texture, and malignancy likelihood. For nodules annotated by multiple radiologists, the dataset provides both individual and consensus-level annotations, enabling downstream tasks such as segmentation, detection, and classification.

The original dataset contains 1,308 CT scans from 1,012 patients. After preprocessing and quality control, the final dataset includes 8,290 2D slices from 669 patients, among whom 383 have malignant nodules and 286 have benign nodules. The slice-level distribution consists of 5,074 malignant and 3,216 benign samples.

## 2.2 Preprocessing

The preprocessing pipeline transformed raw DICOM files from the LIDC-IDRI dataset into standardized nodule image patches suitable for this framework. The pipeline comprises four main steps: DICOM conversion and rescaling, consensus-based ROI cropping, metadata processing, and dataset splitting. An overview of the full preprocessing workflow is illustrated in Figure 1.

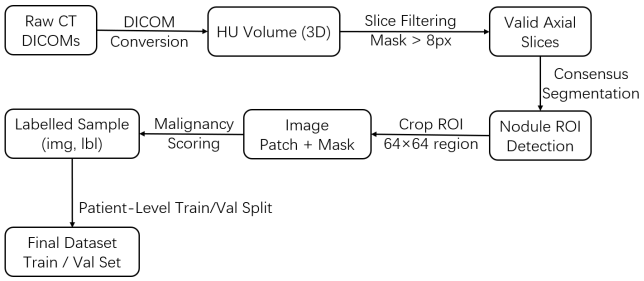



**Figure 1: Preprocessing pipeline for preparing CT lung nodule data for the VAE-MLP framework.**

*2.2.1 DICOM Conversion and Rescaling.* CT scan data is provided in the DICOM format as a sequence of slices. First, read these slices and convert to 3D volumetric array of size 512×512×N, where N is the number of slices per scan. To convert raw DICOM values to medically meaningful Hounsfield Units (HU), we apply the following transformation using DICOM header tags:

$$\text{HU} = (\text{RawValue}) \times \text{RescaleSlope} + \text{RescaleIntercept} \quad (1)$$

To improve data quality and reduce noise, we discard slices with less than 8 pixels in their corresponding segmentation masks. These kinds of images are regarded as noise, not a real nodule, and cannot provide effective training features. Avoid learning irrelevant information from the model.

*2.2.2 Consensus-Based ROI Cropping.* In this dataset, each lung nodule is annotated by four expert radiologists. To improve the quality of the mask, we apply a consensus-based clustering strategy to merge these annotations into a unified segmentation mask. The consensus confidence threshold is set to 0.5. If a pixel appears in at least two radiologists' annotations, it will be considered as part of the nodule. Then, the overlapping parts of the radiologists' annotations are calculated to generate a unified "consensus mask."

Using the final consensus mask, we crop a $64 \times 64$ region of interest (ROI) centered on each nodule. To ensure the cropped ROI is valid and fully within the image, we remove slices where the nodule is too close to the image boundary. This step prevents cropping areas from exceeding the image boundary and ensures the ROI image is complete.

The final ROI image and mask pairs are saved in Numpy files for efficient loading and batching in downstream training.

*2.2.3 Metadata Cleaning and Label Assignment.* Each nodule was also annotated with a malignancy score of 1-5 by four radiologists. The final malignancy score was calculated based on the median score of these four radiologists. Labels are then assigned based on this median:

(1) Malignant (Positive): median score $> 3 \implies$ label = 1
(2) Benign (Negative): median score $< 3 \implies$ label = 0
(3) Uncertain: median score $= 3 \implies$ sample excluded

We exclude ambiguous cases (score = 3) to avoid label noise and improve the model's learning clarity. After this filtering, the dataset size is reduced from 13,607 to 8,290 slices.

*2.2.4 Patient-Level Dataset Splitting.* To ensure that all CTs of the same patient are included into the same set, we split the dataset by patient ID, not by slice. A total of 669 patients is divided into training and validation sets in an 80/20 ratio, ensuring there is no patient overlap between splits. The distribution of the dataset is shown in Figure 2.

The training set is exclusively used for model training, hyperparameter tuning, and classification threshold selection. The validation set is strictly reserved for independent performance evaluation and is never used to adjust or fine-tune any model components. This strict separation ensures that evaluation metrics reflect true generalization.

## 2.3 VAE Training

To help learn meaningful latent representations from lung nodule CT images, we employ a Variational Autoencoder (VAE) as a training module. In this work, all experiments are conducted on 2D slices extracted from 3D CT scans, rather than using full volumetric data. The goal of this stage is to encode high-dimensional 2D image

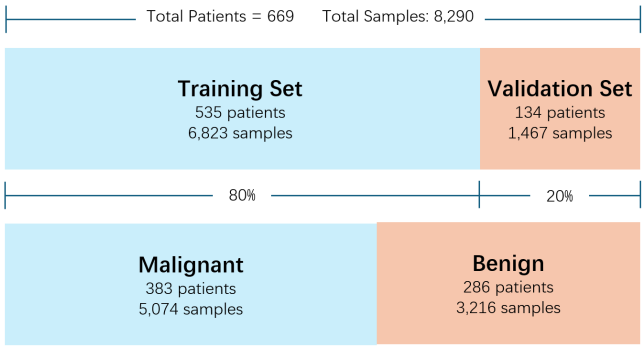



**Figure 2: Dataset split of 669 patients into training (80%) and validation (20%) sets without overlap.**

data into a compact, structured latent space that captures relevant features for subsequent classification and reconstruction tasks.

*2.3.1 VAE Architecture.* To ensure that all sections of the VAE can be viewed as two coupled, but independently parameterized models: the encoder or recognition model, and the decoder or generative model. These two models support each other [2]. The encoder network compresses the input CT image into a latent representation $z$, which parameterizes a Gaussian distribution. The reparameterization trick [3] is used to enable backpropagation through stochastic nodes. It introduces an auxiliary noise variable $\epsilon$, typically sampled from a standard normal distribution, and rewrites the sampling process as $z = \mu + \sigma \cdot \epsilon$, where $\mu$ and $\sigma$ are outputs of the encoder network. This formulation allows gradients to be propagated through the sampling step.

where $\epsilon$ is random noise, $\mu$ is the mean vector of the Gaussian distribution, and $\sigma$ is the logarithm of variance. In this way, gradients can be propagated normally through $\mu$ and $\sigma$, thus achieving end-to-end training.

The decoder network reconstructs the original image from the latent variable $z$ by generating an approximation. Both encoder and decoder are implemented using multi-layer perceptrons (MLPs), with ReLU activation functions and dropout regularization. An overview of the VAE structure used in our framework is shown in Figure 3.

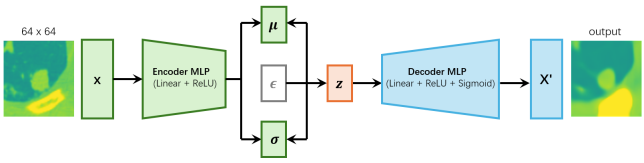



**Figure 3: Overview of the Variational Autoencoder (VAE) structure used in our framework.**

*2.3.2 Hyperparameter Optimization.* To ensure optimal performance, we conduct a systematic hyperparameter search using random exploration. By combining validation loss, SSIM, and MSE, we identify the optimal hyperparameter group. The search space includes the following parameters:

- **HU_LowerBound, HU_UpperBound**: Hounsfield Unit windowing thresholds applied to CT slices.
- **base**: Width scaling factor for the encoder/decoder layers.
- **latent_size**: Dimensionality of the latent space.
- **annealing**: Whether to apply KL annealing during training.
- **alpha, beta**: Loss coefficients for reconstruction and KL divergence terms, respectively.
- **lr**: Learning rate.
- **batch_size**: Batch size used during training.

Each hyperparameter combination is evaluated using a composite score function defined as:

$$\text{Score} = 0.7 \cdot (1 - \text{ValLoss}) + 0.2 \cdot \text{SSIM} + 0.1 \cdot (1 - \text{MSE}) \quad (2)$$

where ValLoss denotes the total validation loss (including KL and reconstruction terms), SSIM measures perceptual similarity, and MSE evaluates pixel-wise reconstruction error. The configuration yielding the highest score is selected for training the final VAE model.

*2.3.3 Output Latent Space.* After training, the encoder produces a latent representation $z$ for each input image. These latent vectors are used as inputs to the downstream MLP classifier and as the initialization for reconstruction refinement. By compressing the CT images into meaningful feature vectors, the VAE helps improve both the diagnostic accuracy and the image reconstruction quality within our framework.

## 2.4 Joint Training with MLP

To enable simultaneous reconstruction and classification from medical image features, we use end-to-end joint fine-tuning that integrates a Variational Autoencoder (VAE) and a Multi-Layer Perceptron (MLP) classifier. After obtaining latent feature vectors from the trained VAE encoder, employing a Multi-Layer Perceptron (MLP) for binary classification of lung nodules. The MLP takes these latent representations as input and outputs a probability indicating whether a nodule is malignant or benign. This joint fine-tuning enables the model to learn both compressed latent representations and discriminative representations for subsequent classification, achieving optimal reconstruction and diagnosis results.

*2.4.1 Model Architecture.* The joint model consists of two main components:

(1) **VAE Encoder-Decoder Network**: This module encodes 3D input volumes into a compact latent space and reconstructs them to enforce representational quality. The encoder produces a latent vector $z$ through a learned Gaussian distribution with mean $\mu$ and log-variance $\log \sigma^2$.
(2) **MLP Classifier**: A lightweight feedforward neural network is used to classify the encoded features. It receives the flattened latent vector $z$ as input and outputs a binary probability.

The MLP classifier consists of a configurable number of fully connected layers. Each layer uses ReLU activation followed by optional dropout regularization. An overview of the joint training architecture is illustrated in Figure 4.

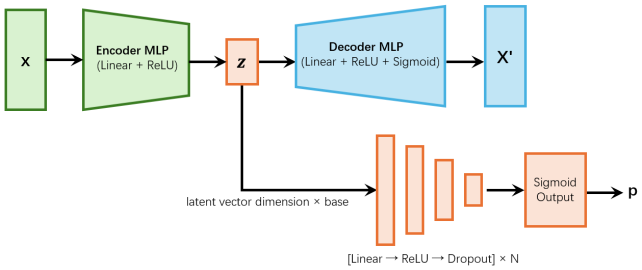



**Figure 4: Architecture diagram of the joint VAE + MLP training framework. The encoder maps CT volumes into latent vectors, which are used for both reconstruction and classification.**

*2.4.2 Loss Component.* In the process of joint fine-tuning, all loss components must be combined to enable the latent space to generate high-quality image reconstructions while retaining strong discriminative power. This enhances both classification performance and the generalization ability of the model. The overall loss function consists of three main parts:

*Reconstruction Loss.* This term measures how well the VAE reconstructs the input image from the latent representation. We use a combination of:

- **L1 Loss**: The mean absolute error between the original image $x$ and the reconstructed image $x'$, which encourages pixel-wise similarity.
- **SSIM Loss**: The Structural Similarity Index measures perceptual quality and structural information, which is important for medical imaging. Since SSIM ranges from 0 to 1 (higher is better), we use $1 - \text{SSIM}$ as the loss term.

The combined reconstruction loss is defined as:

$$\mathcal{L}_{\text{recon}} = \alpha \cdot \text{L1}(x, x') + (1 - \alpha) \cdot (1 - \text{SSIM}(x, x')) \tag{3}$$

where $\alpha$ balances between pixel-level and perceptual similarity.

*Kullback-Leibler Divergence Loss.* This term regularizes the latent space distribution to be close a standard Gaussian prior. This regularization helps the latent space to be smooth, which is essential for good generative performance and prevents overfitting:

$$\mathcal{L}_{\text{KL}} = -\frac{1}{2} \sum (1 + \log \sigma^2 - \mu^2 - \sigma^2) \tag{4}$$

*MLP Classification Loss.* This is a binary cross-entropy (BCE) loss between the predicted label probability $\hat{y}$ and the ground truth label $y$:

$$\mathcal{L}_{\text{MLP}} = -\left[ y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}) \right] \tag{5}$$

*Combined Total Loss.* The total loss optimized at each training step is a weighted sum of the above components:

$$\mathcal{L}_{\text{total}} = \lambda_1 \cdot \mathcal{L}_{\text{recon}} + \lambda_2 \cdot \beta \cdot \mathcal{L}_{\text{KL}} + \lambda_3 \cdot \mathcal{L}_{\text{MLP}} \tag{6}$$

where $\lambda_1$, $\lambda_2$, and $\lambda_3$ control the trade-off between the objectives. In our setting, we choose $\lambda_1 = 0.8$, $\lambda_3 = 1.2$ to slightly prioritize classification accuracy, and $\lambda_2$ and $\beta$ are tuned during training to balance KL divergence regularization.

*2.4.3 Joint Training Strategy.* The joint training framework employs the Adam optimizer with a unified learning rate to simultaneously optimize the parameters of both the Variational Autoencoder (VAE) and the Multi-Layer Perceptron (MLP) sub-models. After loading the preprocessed CT images and their corresponding labels, intensity window normalization was applied within the specified Hounsfield Unit (HU) range to standardize the input. Throughout training, both reconstruction quality metrics, such as Structural Similarity Index Measure (SSIM), and classification performance indicators, including F1-score and Area Under the Curve (AUC), are monitored to evaluate model progress. Upon completion, the best-performing model checkpoint, determined by the lowest validation loss, is restored for final evaluation.

*2.4.4 Evaluation Metrics.* The training procedure follows a 5-fold cross-validation scheme applied exclusively to the training portion of the dataset (Figure 2), ensuring a strict separation between model development and the final evaluation. For each fold, the VAE and MLP models are initialized and jointly fine-tuned to optimize a combined loss function that incorporates both reconstruction and classification objectives. The optimal threshold was found to be 0.7333. Complete each fold metric at Table 2.

An independent performance evaluation is then conducted using the validation set, also shown in Figure 2, which is held out throughout the training and cross-validation process. The final evaluation metrics include precision, recall, F1-score for both benign and malignant classes, as well as ROC-AUC, providing comprehensive insight into model performance. Additionally, a confusion matrix and ROC curve are generated to support further analysis.

## 3 Experiments and Results

### 3.1 Hyperparameter Optimization

To build a high-performing VAE-MLP framework, we conducted extensive hyperparameter searches for both the Variational Autoencoder (VAE) and the downstream Multi-Layer Perceptron (MLP) classifier. This section presents the procedure and results from these searches.

*3.1.1 VAE Hyperparameter Search.* We performed a random search over 100 configurations sampled from a defined hyperparameter space, covering image windowing ranges, architecture complexity, latent space size, KL annealing strategy, and loss weights. Each combination was trained for 10 epochs and evaluated using three metrics: Mean Squared Error (MSE), Structural Similarity Index (SSIM), and Validation Reconstruction Loss.

The VAE hyperparameter space covered a wide range of settings. We explored HU windowing bounds (upper $\in \{300, 400, 500, 600\}$, lower $\in \{-1000, -800, -700, -500\}$), base channel widths $\{32, 64\}$, latent dimensions $\{4, 8, 16\}$, and whether KL divergence was annealed or not. Loss balancing coefficients $\alpha$ and $\beta$ were searched within $[0.15\tilde{}0.85]$ and $[0.5\tilde{}250]$, respectively. The learning rate varied from $1 \times 10^{-6}$ to $1 \times 10^{-4}$, and batch sizes were selected from $\{64, 128\}$.

To identify the optimal configuration for our VAE, we randomly sampled 100 hyperparameter combinations from the defined search space. Across these runs, we observed the best individual metric values as follows: a minimum mean squared error (MSE) of 0.00887,

a maximum structural similarity index (SSIM) of 0.6680, and a minimum validation loss of 0.06157.

We used the scoring function (described in Section 2.3.2) to calculate, the combination of the highest score is the optimal hyperparameter group. Based on this composite score, the highest-performing configuration achieved a score of 0.8621. The corresponding hyperparameter setting included an HU window of $[-1000, 600]$, a base channel width of 64, a latent dimensionality of 4, no KL annealing, $\alpha = 0.8$, $\beta = 5$, a learning rate of $1 \times 10^{-4}$, and a batch size of 128. This best hyperparameter combination was used for all downstream tasks. The top 5 ranking can be viewed in Table 1. The detailed description of each field can be found in Section 2.3.2.

**Table 1: Top 5 VAE hyperparameter configurations ranked by composite score.**

| Score | HU↑ | HU↓ | Base | Latent | KL | $\alpha$ | $\beta$ | LR | Batch |
|---|---|---|---|---|---|---|---|---|---|
| 0.8621 | 600 | -1000 | 64 | 4 | 0 | 0.8 | 5 | 1e-4 | 128 |
| 0.8534 | 300 | -700 | 64 | 4 | 1 | 0.8 | 1 | 1e-4 | 128 |
| 0.8478 | 300 | -1000 | 64 | 8 | 0 | 0.8 | 5 | 5e-5 | 128 |
| 0.8437 | 600 | -800 | 64 | 4 | 0 | 0.7 | 10 | 5e-5 | 128 |
| 0.8372 | 600 | -800 | 64 | 16 | 0 | 0.8 | 10 | 1e-4 | 128 |

*3.1.2 MLP Hyperparameter Search.* For the MLP classifier base on the VAE latent features, we also performed a hyperparameter search of over 50 different combinations. The goal was to maximize the F1 Score on validation set, with the decision threshold fixed at 0.5. The hyperparameter space included the number of hidden layers (depth), layer sizes, and dropout rates. The best performance was achieved by a 4-layer MLP with hidden sizes [2048, 2048, 1024] and a dropout rate of 0.6, yielding the highest validation F1 Score in the search.

## 3.2 Joint Training and Evaluation

After determining the optimal hyperparameters of the VAE and MLP modules, we performed end-to-end joint training to simultaneously optimize image reconstruction and diagnosis. The total training loss consisted of three components: a reconstruction loss that combines L1 and SSIM to ensure both pixel-level accuracy and structural consistency; a Kullback–Leibler (KL) divergence term to regularize the latent space; and a binary cross-entropy loss for nodule classification. To balance reconstruction and classification, we empirically set the loss weights to $\lambda_1 = 0.8$ for reconstruction and $\lambda_3 = 1.2$ for classification, placing slightly more emphasis on diagnostic accuracy.

We evaluated the framework using 5-fold cross-validation on patient-level splits. For each fold, the pre-trained VAE model that achieves the best reconstruction quality was initialized, and the joint model was fine-tuned using the Adam optimizer for 50 epochs. The performance of each fold was monitored using classification metrics including precision, recall, F1-score, and area under the ROC curve (AUC).

Among all five folds, the model trained on fold 5 obtained the best performance both reconstruction and classification. During training, the model produced stable convergence, with reconstruction losses consistently decreasing across epochs. To calibrate the classification output, we evaluated multiple threshold values and selected the one that maximized the F1 score across the test sets of all folds. The optimal threshold was found to be 0.7333. Complete classification metrics for each fold and the test set are shown in Table 2. Performance on the validation set is reported separately in Table 3.

**Table 2: Classification performance metrics under the unified optimal threshold (0.7333). Precision, recall, and F1 scores are reported separately for malignant (positive) and benign (negative) classes across each fold and for the test set.**

| Fold | Malignant (Positive) | | | Benign (Negative) | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 |
| fold0 | 0.9448 | 0.9438 | 0.9443 | 0.9115 | 0.9129 | 0.9122 |
| fold1 | 0.9309 | 0.9429 | 0.9369 | 0.9079 | 0.8896 | 0.8987 |
| fold2 | 0.9364 | 0.9429 | 0.9396 | 0.9088 | 0.8989 | 0.9038 |
| fold3 | 0.9270 | 0.9379 | 0.9324 | 0.9002 | 0.8834 | 0.8917 |
| fold4 | 0.9579 | 0.9428 | 0.9503 | 0.9121 | 0.9348 | 0.9233 |
| **Test** | **0.9394** | **0.9420** | **0.9407** | **0.9081** | **0.9039** | **0.9059** |

**Table 3: Classification metrics on the validation set under the same optimal threshold (0.7333).**

| Set | Malignant (Positive) | | | Benign (Negative) | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 |
| **Valid** | **0.9885** | **0.9916** | **0.9900** | **0.9844** | **0.9787** | **0.9816** |

Under this threshold, the model was evaluated on the validation set achieving an F1 score of 0.9503, indicating high classification reliability. The precision and recall for the malignant class were 0.9579 and 0.9428, respectively, while the AUC reached 0.9782, reflecting excellent overall separability. The confusion matrix further demonstrated the model's robust generalization, with only 42 false positives and 58 false negatives in the validation set. These results confirm that the joint optimization strategy enabled the latent space to be simultaneously informative for both image reconstruction and diagnostic classification.

## 4 Conclusion

In this paper, we proposed a VAE-MLP framework for lung nodule reconstruction and diagnosis from CT images. By combining the generative power of Variational Autoencoders with the discriminative ability of Multi-Layer Perceptron, our method effectively captures latent features that support both high-fidelity reconstruction and accurate malignancy classification.

Many experiments performed on the LIDC-IDRI dataset show that our framework achieves high diagnostic performance, with best F1-scores and ROC-AUC, and preserve perceptual image quality. The joint fine-tuning helps with mutual enhancement between the encoder-decoder and classifier, making the framework robust and interpretable. In general, our work shows the potential of hybrid generative-discriminative models in medical imaging and provides promising direction for future research in automated diagnosis systems.

However, there are some limitations to our study. First, our model was trained and tested on 2D slices rather than full 3D volumes, which might miss important spatial and contextual information inherent in volumetric data. Second, we excluded nodules with a malignancy score of 3 to reduce label noise, but this also reduced dataset diversity. Additionally, our approach relies solely on CT imaging data without incorporating other potentially informative modalities.

For future work, we plan to explore 3D convolutional architectures to better capture volumetric context and spatial relationships. We also aim to incorporate semi-supervised learning techniques to leverage uncertain or weakly labeled data more effectively. Furthermore, integrating multi-modal data sources such as PET scans or clinical information could enhance the diagnostic performance and robustness of the model.

## References

[1] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639 (Feb. 2017), 115–118.

[2] Diederik P. Kingma and Max Welling. 2019. An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning* 12, 4 (2019), 307–392. doi:10.1561/2200000056

[3] Diederik P Kingma and Max Welling. 2022. Auto-Encoding Variational Bayes. arXiv:1312.6114 [stat.ML] https://arxiv.org/abs/1312.6114

[4] Chao Li, Shaoyuan Lei, Li Ding, Yan Xu, Xiaonan Wu, Hui Wang, Zijin Zhang, Ting Gao, Yongqiang Zhang, and Lin Li. 2023. Global burden and trends of lung cancer incidence and mortality. *Chin Med J (Engl)* 136, 13 (July 2023), 1583–1590.

[5] D R Sarvamangala and Raghavendra V Kulkarni. 2021. Convolutional neural networks in medical image understanding: a survey. *Evol Intell* 15, 1 (Jan. 2021), 1–22.

[6] Durgesh Srivastava, Santosh Kumar Srivastava, Surbhi Bhatia Khan, Hare Ram Singh, Sunil K Maakar, Ambuj Kumar Agarwal, Areej A Malibari, and Eid Albalawi. 2023. Early Detection of Lung Nodules Using a Revolutionized Deep Learning Model. *Diagnostics (Basel)* 13, 22 (Nov. 2023).

[7] Rui Wei and Abdul Mahmood. 2021. Recent Advances in Variational Autoencoders with Representation Learning for Biomedical Informatics: A Survey. *IEEE Access* 9 (2021), 4939–4956. doi:10.1109/ACCESS.2020.3048309

[8] World Health Organization, International Agency for Research on Cancer (IARC). n.d.. Lung Cancer. https://www.iarc.who.int/cancer-type/lung-cancer/. Accessed: 2025-07-01.