# Explainable COVID-19 Classification Via Variational and Perceptual Autoencoder-Guided Occlusion

**Rodrigo Bayuk, Joel Manquel, Orietta Nicolis, Billy Peralta**
Universidad Andres Bello, Facultad de Ingeniería,
Santiago, Chile, 11300,
{*r.bayuknunez,j.manquelmanquel*}*@uandresbello.edu,*{*orietta.nicolis,billy.peralta*}*@unab.cl*

and

**Luis Caro**
Universidad Católica de Temuco, Departamento de Ingeniería Civil en Informática,
Temuco, Chile, 11300,
*lcaro@uct.cl*

**Abstract**

Technology played a crucial role in combating the COVID pandemic, both in the rapid development of vaccines and the early detection of the virus. Consequently, numerous studies in the medical field have focused on leveraging the power of artificial intelligence for COVID-19 detection. However, in the medical domain, it is essential to have a clear understanding of the processes and algorithms used in decision-making, as these directly impact people's health. Therefore, efforts have been made to implement explainable artificial intelligence techniques, enabling humans to understand and explain the deep learning algorithms used in disease detection. In this work, we present an approach to detecting COVID-19 in chest X-rays that combines reconstruction-based anomaly discovery with perturbation-based attribution. Specifically, we use a Variational Autoencoder trained on healthy lungs to identify lung anomalies, and we additionally evaluate a Perceptual Autoencoder (PAE) that incorporates a perceptual loss to produce sharper reconstructions and more contrastive residual maps. These signals are then used to highlight critical image evidence through both patch-level scoring (grid-based occlusion) and pixel-level masking derived from reconstruction-error maps, enabling healthcare professionals to localize relevant regions more effectively. Moreover, the proposed framework provides clearer explanations of the model's decisions by quantifying the prediction change after occluding the identified regions, reducing the complexity of the "black boxes" generated by deep learning neural networks. With this methodology, we aim to improve the effectiveness and reliability of early COVID-19 detection through chest X-rays.

**Keywords:** COVID classification, Artificial Vision, Variational autoncoder, Perceptual autoncoder.

## 1 Introduction

Medical image interpretation through chest radiographs has become a fundamental tool for rapid diagnosis and treatment planning in clinical settings [1]. During the COVID-19 pandemic, the overwhelming number of cases placed unprecedented pressure on healthcare systems worldwide, underscoring the need for diagnostic tools that are both accurate and time-efficient [2]. In Chile, for example, gold-standard RT-PCR tests often required more than 24 hours to deliver results, contributing to delays in patient management and increased strain on intensive care units [3]. Chest radiography, a non-invasive imaging technique, captures detailed views of the lungs and heart and plays a crucial role in diagnosing a variety of thoracic conditions. In the context of suspected COVID-19, chest X-rays can provide valuable insights into pulmonary involvement, especially in resource-limited or high-demand settings [4]. To support clinical interpretation, the Dutch Society of Radiology introduced a standardized protocol for chest CT evaluation in COVID-19 cases, known as CO-RADS. This framework categorizes radiological findings into five levels of suspicion, ranging from

very low (CO-RADS 1) to very high (CO-RADS 5), and is designed for use in patients with moderate to severe symptoms in regions with significant disease prevalence [4].

Recent advances in deep learning, particularly convolutional neural networks (CNNs) such as ResNet variants, have achieved impressive performance—exceeding 95% accuracy—in detecting COVID-19 from chest X-ray images [5]. Nevertheless, the reliance on opaque, "black-box" models remains a critical limitation. Without a clear rationale behind a model's prediction, clinicians are often reluctant to incorporate automated results into high-stakes decision-making. Traditional CNNs excel at learning complex feature hierarchies from large datasets, but they often ignore domain-specific constraints—such as anatomical localization of lung opacities—that experts deem essential [5]. In practice, models may latch onto spurious artifacts (e.g., markers or noise outside the lungs) that correlate with disease labels rather than true pathological patterns. Consequently, the problem of explainability (i.e., providing human-interpretable evidence for a given prediction) remains unresolved in most AI pipelines for radiology. Existing visualization methods like GradCAM highlight regions of activation, but they do not guarantee correlation with medically relevant abnormalities [6]. Moreover, there is a lack of standardized frameworks to integrate domain knowledge (e.g., radiological criteria) into model training and inference. This gap motivates the need for new architectures and loss functions that explicitly enforce attention within medically plausible areas, so that AI decisions align with expert reasoning and build clinical trust.

In this work, we propose a framework based on a Variational Autoencoder (VAE) for explainable COVID-19 classification from chest X-rays. First, we train the VAE exclusively on healthy lung images to learn a probabilistic latent representation of normal pulmonary tissue. During inference, any COVID-19 image produces higher reconstruction error in the pathological regions, automatically highlighting pulmonary anomalies. We employ a VAE loss to balance reconstruction fidelity and latent-space regularization, ensuring that the learned features capture clinically relevant variations. From the "virtually healthy" reconstruction generated by the decoder, we compute anomaly maps that precisely indicate the affected areas. In addition, we evaluate a Perceptual Autoencoder (PAE) trained on Normal images with a perceptual loss term, which promotes sharper reconstructions and more contrastive residual maps for highlighting key evidence. Finally, we use these anomaly-derived signals to highlight critical regions through both patch-based occlusion (grid scoring) and pixel-level masking, and a secondary classifier uses the resulting perturbations to assign and validate a COVID-19 score. This joint reconstruction–classification strategy enables both accurate detection and interpretable localization of lesions, since the encoder–decoder explicitly exposes deviations from the normal lung pattern.

This article extends our preliminary conference version [7], which introduced VAE-guided patch occlusion for explainable COVID-19 classification. The present version substantially expands that work by incorporating a Perceptual Autoencoder with perceptual loss, adding pixel-wise reconstruction-error masking, including a GradCAM baseline, extending the quantitative evaluation, and providing additional qualitative analyses.

## 2 Related Works

Deep learning has played a central role in advancing medical image analysis, particularly in the classification and detection of thoracic diseases using chest X-rays (CXRs). Traditional convolutional neural networks (CNNs) and more recent approaches leveraging variational autoencoders (VAEs), perceptual autoencoders (PAEs), transfer learning, and explainable AI have all been explored to enhance diagnostic accuracy and interpretability.

Several works have focused on improving multi-label classification of chest pathologies. Allaouzi and Ahmed [8] proposed a novel approach for multi-label classification of thorax diseases using CXR images. More recently, Hanif et al. [9] introduced an improved ranking loss function that significantly enhanced multi-label classification performance. Zhao and Wang [10] proposed a model that captures long-range dependencies and label relationships, further improving classification results on CXR datasets.

Explainability has become a critical feature in clinical AI applications. Chetoui et al. [11] proposed an end-to-end explainable CNN for COVID-19 detection, offering clinicians transparent insights into decision-making. Rocha et al. [12] presented CLARE-XR, a regression-based explainable model that incorporates label embeddings to improve interpretation of predictions.

Unsupervised approaches using VAEs have also been explored. Mansour et al. [13] and Addo et al. [14] developed VAE-based and ensemble VAE-based frameworks, respectively, to detect COVID-19 from CXR images, demonstrating the potential of generative models for unsupervised anomaly detection. Similarly, Zhou et al. [15] introduced WVALE, a weakly supervised VAE for localizing and enhancing lung infection regions in COVID-19 patients. Jahan and Hasan [16] proposed an autoencoder-based anomaly detector specifically tailored for COVID-19 screening in CXRs.

Federated and distributed learning frameworks have also emerged to address data privacy concerns. Guan

et al. [17] presented a comprehensive survey of federated learning strategies in medical imaging, outlining architectures and security trade-offs relevant to multi-institutional data analysis.

In terms of deep learning developments, Chen et al. [18] reviewed the clinical applicability of advanced CNNs, VAEs, and transformer-based models in medical image analysis, emphasizing their utility in real-world diagnostics. Martinez et al. [19] proposed a deep unsupervised method based on patch-wise analysis for identifying relevant regions in X-rays, showing promising results in localized COVID-19 detection.

These contributions collectively highlight the importance of combining high-performance classification with explainability and localization, especially when addressing urgent public health needs such as COVID-19 detection from chest X-rays. The main contribution of our work in relation to the existing literature is the combination of explainability and precision in COVID-19 detection from chest X-rays through an innovative integration of Variational Autoencoders (VAEs) and GradCAM-based occlusion methods, further strengthened by evaluating a Perceptual Autoencoder (PAE) to produce sharper reconstructions and more contrastive residual maps. By unifying reconstruction based anomaly localization with perturbation-based attribution, our framework highlights clinically relevant evidence at multiple granularities: coarse *patch-level* importance via grid scoring and fine *pixel-level* masking derived from reconstruction errors. This design not only improves the spatial specificity of the highlighted regions compared to activation-only visualizations, but also provides a quantitative sanity check by measuring the prediction drop after occluding the identified areas, thereby offering a more faithful and clinically actionable explanation of the classifier's decisions. Compared with the previous conference paper [7], this version introduces new methodological and experimental material, including the PAE-based explanation strategy, pixel-level masking, additional baselines, extended metrics, and a more detailed analysis of explanation fidelity.

## 3 Theoretical Background

In this section, we outline the core components underpinning our explainable COVID-19 classification framework. We begin by reviewing DenseNet [20] which serves as robust feature extractor for chest X-ray images. To localize class-relevant regions, we employ GradCAM [6] to generate attention heatmaps that guide interpretability. Finally, we introduce Autoencoders, with a particular focus on Variational Autoencoders (VAEs) [21] and Perceptual Autoencoders (PAEs) [22], which learn a latent representation of healthy lung tissue from Normal chest X-rays; during inference, reconstruction errors highlight anomalies associated with COVID-19, while the perceptual loss in the PAE promotes sharper reconstructions and more contrastive residual maps for lesion highlighting.

### 3.1 Densely Connected Networks (DenseNet)

DenseNets [20] further improve gradient flow and parameter efficiency by connecting every layer to all subsequent layers in a feedforward manner. In a DenseNet, the $\ell$-th layer receives as input the feature-maps of all preceding layers:

$$x_\ell = H_\ell\big([x_0, x_1, \ldots, x_{\ell-1}]\big),$$

where $[x_0, x_1, \ldots, x_{\ell-1}]$ denotes the concatenation of feature-maps from layers 0 through $\ell - 1$, and $H_\ell(\cdot)$ represents a composite of batch normalization, ReLU, and convolution. Unlike ResNet's additive identity skips, DenseNet's concatenation enforces that each layer has direct access to all previously learned feature representations. This dense connectivity pattern encourages feature reuse, reduces the number of parameters (since each layer only produces a small growth in feature channels), and ensures robust gradient propagation during backpropagation. As a result, DenseNets often match or surpass traditional CNNs with fewer parameters, offering both computational and memory efficiency.

### 3.2 Autoencoders

An Autoencoder is an unsupervised neural architecture designed to learn compact latent representations of input data through a two-stage process: encoding and decoding. The **encoder** maps an input $\mathbf{x} \in \mathbb{R}^n$ to a lower-dimensional latent vector $\mathbf{z} \in \mathbb{R}^d$ $(d < n)$, effectively filtering out irrelevant features. Formally:

$$\mathbf{z} = f_{\text{enc}}(\mathbf{x}; \theta_{\text{enc}}),$$

where $f_{\text{enc}}$ is typically a stack of convolutional or fully connected layers. The **decoder** then reconstructs $\hat{\mathbf{x}}$ from $\mathbf{z}$:

$$\hat{\mathbf{x}} = f_{\text{dec}}(\mathbf{z}; \theta_{\text{dec}}),$$

so that $\hat{\mathbf{x}} \approx \mathbf{x}$. Training minimizes a reconstruction loss (e.g., mean-squared error) between $\mathbf{x}$ and $\hat{\mathbf{x}}$. Variants include the *Denoising Autoencoder*, which learns to reconstruct clean inputs from corrupted versions, and

the *Sparse Autoencoder*, which enforces sparsity in $\mathbf{z}$. The basic autoencoder framework underpins many applications, from dimensionality reduction to anomaly detection and generative modeling.

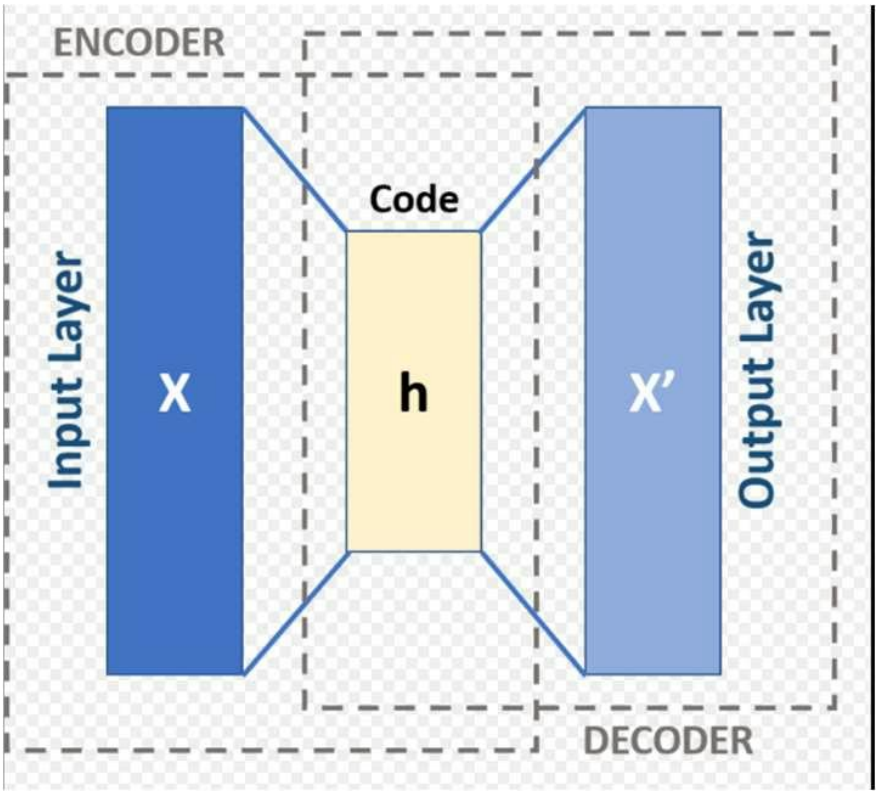


Figure 1: Generic Autoencoder Architecture

### *3.2.1 Variational Autoencoder (VAE)*

Variational Autoencoders (VAEs) [21] extend the autoencoder paradigm by learning a probabilistic latent space rather than a deterministic code. Instead of mapping an input $\mathbf{x}$ to a single latent point, the encoder produces the parameters of an approximate posterior distribution $q_\phi(\mathbf{z} \mid \mathbf{x})$, typically a diagonal Gaussian with mean and variance:

$$q_\phi(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}\big(\mathbf{z};\, \boldsymbol{\mu}_\phi(\mathbf{x}), \mathrm{diag}(\boldsymbol{\sigma}^2_\phi(\mathbf{x}))\big).$$

Sampling $\mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{x})$ and passing it through the decoder $p_\theta(\mathbf{x} \mid \mathbf{z})$ yields a reconstruction. In practice, sampling is implemented via the reparameterization trick:

$$\mathbf{z} = \boldsymbol{\mu}_\phi(\mathbf{x}) + \boldsymbol{\sigma}_\phi(\mathbf{x}) \odot \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

which enables backpropagation through stochastic latent variables. Training maximizes the evidence lower bound (ELBO):

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] \;-\; D_{\mathrm{KL}}\big(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z})\big),$$

where $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the prior. The first term encourages accurate reconstruction (e.g., via a Gaussian or Bernoulli likelihood), while the Kullback–Leibler term regularizes the latent space by keeping the learned posterior close to the prior, promoting smoothness and improving generalization. For the common Gaussian posterior and standard normal prior, the KL term has a closed form:

$$D_{\mathrm{KL}}\big(\mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2)) \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I})\big) = \frac{1}{2} \sum_j \left(\mu_j^2 + \sigma_j^2 - \log \sigma_j^2 - 1\right).$$

VAEs generate new samples by drawing $\mathbf{z} \sim p(\mathbf{z})$ and decoding to $\mathbf{x}$. This probabilistic foundation makes VAEs a cornerstone for tasks requiring both representation learning and generative capabilities, while also enabling anomaly sensitive behavior when inputs deviate from the learned data manifold.

In our implementation, we use a $\beta$-VAE formulation, where the KL regularization term is weighted by a fixed hyperparameter $\beta$:

$$\beta D_{\mathrm{KL}}\big(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z})\big)$$

Here, $\beta$ controls the relative strength of latent-space regularization with respect to reconstruction fidelity.

#### *3.2.2 Perceptual Autoencoder (PAE)*

Perceptual Autoencoders (PAEs) [23, 22] extend the standard autoencoder paradigm by augmenting pixel-level reconstruction with a *perceptual similarity* objective, encouraging reconstructions that are not only structurally accurate but also *sharp* and *texture-consistent*. Given an input image $\mathbf{x}$, a deterministic encoder $f_\phi(\cdot)$ maps it to a latent representation $\mathbf{z}$,

$$\mathbf{z} = f_\phi(\mathbf{x}),$$

which is then decoded by $g_\theta(\cdot)$ to produce a reconstruction,

$$\hat{\mathbf{x}} = g_\theta(\mathbf{z}) = g_\theta\big(f_\phi(\mathbf{x})\big).$$

Unlike VAEs, PAEs do not impose a probabilistic prior over $\mathbf{z}$ and therefore do not include a KL regularization term; the latent code is optimized purely through reconstruction objectives. The defining component is the use of a perceptual distance computed in the feature space of a pretrained network. A common instantiation is LPIPS [22], which measures learned perceptual similarity via deep features. Denoting this term by $\mathcal{L}_{\mathrm{LPIPS}}(\mathbf{x}, \hat{\mathbf{x}})$, the PAE is trained with:

$$\mathcal{L}_{\mathrm{total}}(\theta, \phi; \mathbf{x}) = \mathcal{L}_1(\mathbf{x}, \hat{\mathbf{x}}) + \lambda\, \mathcal{L}_{\mathrm{LPIPS}}(\mathbf{x}, \hat{\mathbf{x}}),$$

where $\mathcal{L}_1$ enforces global anatomical structure while the perceptual component emphasizes high-frequency details and visually plausible texture. In practice, the perceptual term can be written as a weighted sum of feature distances across multiple layers $\ell$ of a fixed feature extractor $\psi$:

$$\mathcal{L}_{\mathrm{LPIPS}}(\mathbf{x}, \hat{\mathbf{x}}) = \sum_\ell w_\ell \left\| \psi_\ell(\mathbf{x}) - \psi_\ell(\hat{\mathbf{x}}) \right\|_2^2,$$

with learned or predefined weights $w_\ell$ depending on the LPIPS variant. By construction, PAEs tend to yield crisper reconstructions than purely pixel based autoencoders, which is beneficial when reconstruction residuals are later used as an anomaly signal. When trained on Normal chest X-rays, a PAE learns a high-fidelity manifold of healthy appearance; during inference, pathological regions that cannot be faithfully represented under this Normal prior produce larger residuals, enabling anomaly-sensitive behavior analogous to VAE-based approaches but without probabilistic latent regularization.

## 4 Proposed Methodology

Accurate COVID-19 detection in chest X-rays demands not only high classification performance but also transparent localization of the underlying anomalies. In this proposal, a Variational AutoEncoder (VAE) architecture was employed and trained exclusively with chest X-ray images of healthy lungs. The goal was to generate only normal images, thereby enabling more accurate detection of COVID-related anomalies. However, during the reconstruction tests, it was observed that the accuracy in generating normal images was so high that, when presented with a COVID-19 image, the neural network was capable of reconstructing it with a high level of detail—rather than producing a normal-looking image.

This outcome is explained by the fact that traditional autoencoders learn a fixed representation of the input data. Since the features of COVID-affected lungs can be visually similar to those of healthy lungs, the model was able to reconstruct the original image completely—even when the input came from an infected patient.

To improve the precision in detecting COVID-19 images, a different training approach was adopted: the use of a Variational AutoEncoder. This model introduces variability into the image generation process by learning a probability distribution over possible representations, rather than a fixed encoding. In this way, the model approximates an unknown distribution of latent features, enabling the detection of deviations more effectively.

Moreover, the loss function included a regularization parameter called $\beta$ (beta), designed to control the flexibility of the latent space representation. The inclusion of this parameter in the VAE resulted in a slight reduction in reconstruction accuracy, but it improved the conformity of reconstructed COVID-related images to the normal distribution, which in turn enhanced the ability to detect such anomalies.

In addition, we evaluate a Perceptual Autoencoder (PAE) trained exclusively on Normal images, which complements the VAE by incorporating a perceptual term in the reconstruction objective (e.g., $L_{\mathrm{total}} =$

$L_1 + \lambda L_{\text{LPIPS}}$). This perceptual constraint encourages sharper, more photorealistic reconstructions with better high-frequency texture consistency; consequently, when a COVID-19 image is reconstructed using only Normal priors, residual maps tend to become more contrastive in lesion-like regions, improving the highlighting of key evidence.

At inference, any input exhibiting pathological patterns, such as COVID-19 related opacities, yields elevated pixel-wise reconstruction error, thereby providing a principled, unsupervised anomaly map without the need for lesion annotations.

As a comparative baseline, we also propose a patch-based occlusion scheme: the X-ray is divided into equal-sized patches, each is independently scored by a pretrained DenseNet201 classifier, and the highest-scoring region is masked to measure its impact on the model's output. Detailed descriptions of these approaches are provided in Sections 4.1 (Patch-Based Occlusion), 4.2 (VAE-Based Occlusion), and 4.3 (PAE-based occlusion).

### 4.1 Patch-Based Occlusion (4x4 Grid)

Patch-based occlusion is a simple, model-agnostic perturbation strategy to estimate *regional importance* for a COVID-19 classifier. The key idea is to (i) decompose the input chest X-ray into coarse spatial regions, (ii) measure how confident the classifier is when it sees each region in isolation, and (iii) occlude the most "COVID-evidential" region in the full image and quantify the resulting drop in the predicted COVID-19 score. This produces an intuitive attribution signal: if removing a region causes a large confidence decrease, that region likely contains discriminative evidence.

The **default configuration** uses a 4×4 partition (16 non-overlapping patches). The same procedure naturally generalizes to any $m \times n$ grid resolution; in our experiments, we additionally evaluate a 3×3 partition (9 patches) to study the effect of coarser spatial granularity on the estimated attribution.

Let $p_{\text{orig}} = \text{DenseNet201}(x)$ denote the baseline COVID-19 score on the original input image $x$. We then proceed as follows:

1. **Patch Extraction.** Divide the input image $x$ into sixteen non-overlapping patches $\{P_i\}_{i=1}^{16}$. This coarse partition enforces locality while remaining computationally lightweight.
2. **Patch Classification.** Compute a COVID-19 score for each patch by feeding it independently to the same backbone:
$$p_i = \text{DenseNet201}(P_i).$$
Intuitively, patches that contain pathological patterns (e.g., opacities) should yield higher $p_i$.
3. **Critical Patch Selection.** Select the most influential patch index as the one maximizing the COVID-19 score:
$$k = \arg\max_i p_i.$$
This patch is treated as the *most critical region* under the patch-wise evidence criterion.
4. **Occlusion.** Mask all pixels in $P_k$ within the full image by filling them with zero to obtain $\tilde{x}$.
5. **Reclassification.** Recompute the model score on the occluded image:
$$p_{\text{occ}} = \text{DenseNet201}(\tilde{x}).$$
6. **Impact Quantification.** Quantify the relative impact of removing the critical region as:
$$\Delta p = \frac{p_{\text{orig}} - p_{\text{occ}}}{p_{\text{orig}}} \times 100\%.$$
Larger $\Delta p$ indicates a stronger dependence of the prediction on the selected patch.

### 4.2 VAE-Based Occlusion (4x4 Grid)

VAE-based occlusion leverages *reconstruction error* as an unsupervised proxy for abnormality localization. A variational autoencoder (VAE) is trained *only* on healthy (Normal) chest X-rays, thus learning a compact distribution of normal pulmonary appearance. At inference time, when presented with a potentially COVID-19 image, the VAE tends to reconstruct a "healthy-looking" counterpart; consequently, pathological regions are not well reconstructed and yield high pixel-wise reconstruction residuals. We aggregate this residual signal over a 4×4 grid to identify the most anomalous patch, occlude it in the original image, and measure the corresponding drop in the DenseNet201 COVID-19 score. This pipeline provides a classifier-agnostic, reconstruction-driven attribution signal.

The **default configuration** uses a 4×4 partition (16 non-overlapping patches). The exact procedure is:

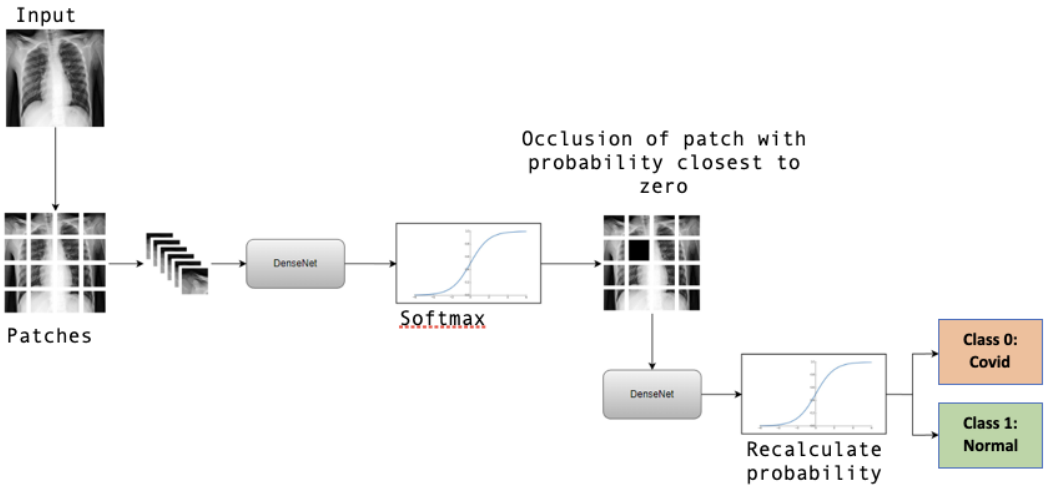


Figure 2: Patch-based occlusion using a $4 \times 4$ grid.

1. **Baseline Classification.** Compute $p_{\text{orig}} = \text{DenseNet201}(x)$.
2. **VAE Reconstruction.** Generate $x' = \text{VAE}(x)$.
3. **Error Mapping.** Form the absolute difference map:
$$\Delta(u, v) = |x(u, v) - x'(u, v)|.$$
4. **Patch Scoring.** Split $\Delta$ into 16 non-overlapping patches $\{P_i\}_{i=1}^{16}$ according to a $4\times4$ grid, and compute
$$S_i = \sum_{(u,v)\in P_i} \Delta(u, v).$$
5. **Critical Patch Selection.** Select the most anomalous patch:
$$k = \arg\max_i S_i.$$
6. **Occlusion and Reclassification.** Mask all pixels in $P_k$ within the original image $x$ by filling them with zero to obtain $\tilde{x}$, then evaluate
$$p_{\text{occ}} = \text{DenseNet201}(\tilde{x}).$$
7. **Impact Quantification.** Compute
$$\Delta p = \frac{p_{\text{orig}} - p_{\text{occ}}}{p_{\text{orig}}} \times 100\%.$$

### 4.3 PAE-Based Occlusion (Perceptual Loss)

PAE-based occlusion follows the same reconstruction-error attribution principle as Sec. 4.2, but replaces the VAE with a **Perceptual Autoencoder (PAE)** to prioritize *sharper, more photorealistic* reconstructions via a perceptual term. The PAE is a convolutional encoder–decoder with a strict bottleneck and no long-range skip connections (to avoid copying anomalies through shortcuts). The decoder produces outputs through a `tanh` activation, i.e., reconstructions lie in $[-1, 1]$. At inference time, the PAE reconstructs a COVID-19 input using only Normal priors; lesion regions are typically "restored" toward healthy appearance and therefore yield high absolute residuals. Unlike the VAE variant that selects a single critical patch, here the residual is thresholded into an adaptive *pixel-wise* occlusion mask, which is applied to the original image to perturb the input prior to DenseNet201 reclassification. We choose this option to compare an alternative way to oclude input image. The exact procedure is:

1. **Baseline Classification.** Compute $p_{\text{orig}} = \text{DenseNet201}(x)$.

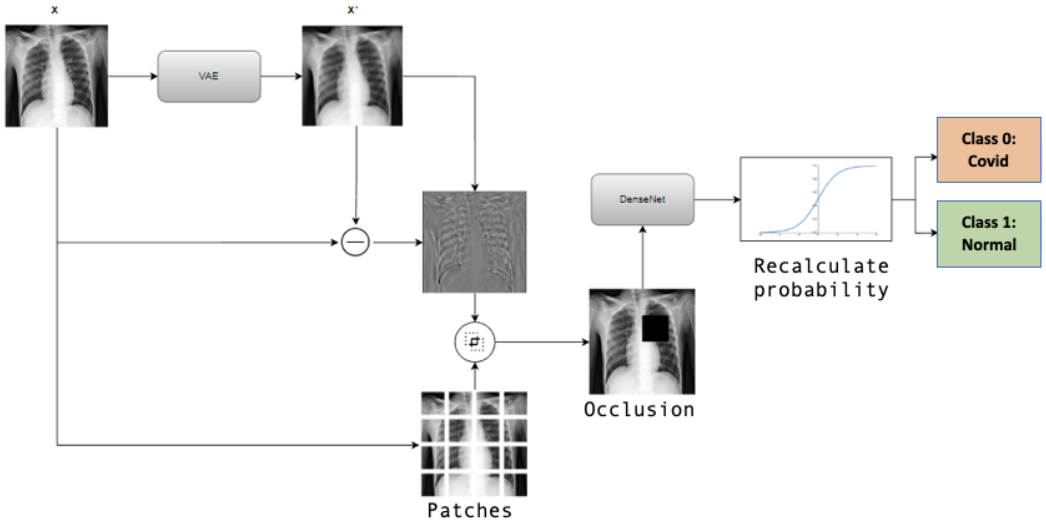


Figure 3: VAE-based occlusion pipeline using a $4 \times 4$ grid.

2. **PAE Reconstruction.** Generate $\hat{x} = \text{PAE}(x)$. (Both $x$ and $\hat{x}$ are assumed to be in the same normalized intensity space; in our implementation $\hat{x} \in [-1, 1]$).

3. **Error Mapping.** Form the absolute reconstruction error map:

$$\Delta(u, v) = |x(u, v) - \hat{x}(u, v)|.$$

4. **Adaptive Thresholding.** Compute the threshold

$$\tau = \mu_\Delta + 0.5\,\sigma_\Delta,$$

where $\mu_\Delta$ and $\sigma_\Delta$ are the mean and standard deviation of $\Delta$.

5. **Mask Construction.** Build the occlusion mask

$$M(u, v) = \mathbb{I}\big[\Delta(u, v) > \tau\big].$$

6. **Occlusion and Reclassification.** Apply the mask to the original image to obtain $\tilde{x}$ by filling masked pixels with zero (i.e., $\tilde{x}(u, v) = 0$ if $M(u, v) = 1$), then evaluate

$$p_{\text{occ}} = \text{DenseNet201}(\tilde{x}).$$

7. **Impact Quantification.** Compute

$$\Delta p = \frac{p_{\text{orig}} - p_{\text{occ}}}{p_{\text{orig}}} \times 100\%.$$

## 5 Data analysis

### 5.1 Dataset

For this study, we utilized the COVID-19 Radiography Database, a comprehensive collection of chest X-ray images containing confirmed COVID-19 cases, normal (healthy) lungs, and viral pneumonia instances. This dataset was obtained from Kaggle and was originally compiled by researchers at Qatar University and the University of Dhaka, in collaboration with colleagues from Pakistan and Malaysia, along with medical professionals [24].

These images present a variety of radiographic patterns—such as ground-glass opacities, consolidation, and normal lung fields—that serve as the primary input for all neural network experiments. A high-quality, large-scale dataset is critical for training robust deep learning models, and the COVID-19 Radiography Database provides sufficient heterogeneity in terms of disease severity, patient demographics, and imaging equipment variability to ensure reliable performance and generalization.

Specifically, the database comprises 11,000 images labeled as "Normal," 3,700 images labeled as "COVID-19," 5,700 images labeled as "Opacity," and 1,500 images labeled as "Pneumonia." The balanced distribution of these classes allows our model to learn discriminative features across healthy and pathological lung tissues, which is essential for accurate and explainable COVID-19 classification.

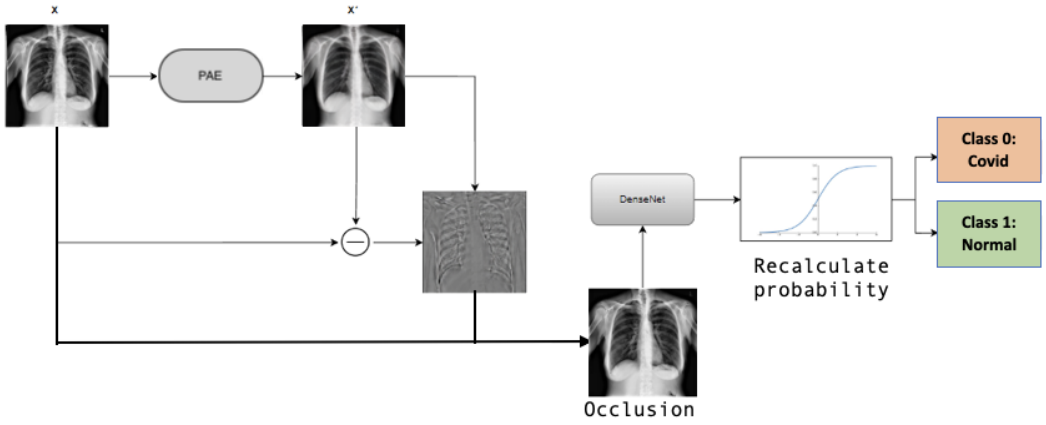


Figure 4: PAE-based occlusion pipeline.

### 5.2 Data Characteristics

High-quality imaging data is essential for the reliable training of machine learning models. In our chest X-ray dataset, we encountered several quality-related challenges that can adversely affect model performance. Specifically, aside from low resolution, many images exhibit artifacts, noise, and unwanted elements that must be addressed during preprocessing. Careful curation and filtering of these cases are necessary to ensure that the neural network learns from diagnostically relevant features. Below are representative issues detected in the raw dataset:

- **Incomplete Framing:** Radiographs that do not occupy the full image frame, leaving large blank margins or cropping out critical lung regions.
- **Superimposed Annotations:** Added arrows, text labels, or markers that obscure anatomical structures.
- **Suboptimal Acquisition Quality:** Images degraded at source due to poor X-ray technique, resulting in under- or overexposure.
- **Noise and Compression Artifacts:** High levels of sensor noise or JPEG compression artifacts that obscure fine-grained pulmonary details.
- **Extraneous Objects in Frame:** Foreign objects (e.g., tubes, electrodes) unintentionally captured within the X-ray field.
- **Internal Implants and Prostheses:** Presence of pacemakers, surgical clips, or prosthetic devices that introduce high-contrast shadows.
- **Surgical or Interventional Artifacts:** Visual artifacts from recent procedures, such as chest drains or stents, which may confound pathology detection.
- **Duplicate or Near-Duplicate Images:** Redundant images that can bias the training set if not properly deduplicated.

Addressing these data quality issues through manual inspection, automated filtering, or augmentation strategies, is crucial to maintain diagnostic fidelity and to train a robust, generalizable model.

#### *5.2.1 Interpretation of Chest X-Ray Intensity Values*

All analyzed radiographs are stored as grayscale images, where pixel intensities range from 0 to 255: an intensity of 0 corresponds to pure black (maximum radiolucency), while 255 represents pure white (maximum radiopacity). In chest X-ray imaging, different anatomical structures and pathologies manifest as characteristic intensity ranges:

- **Air-filled regions (e.g., healthy lung fields):** Appear nearly black due to low density and high radiolucency.
- **Subcutaneous fat:** Renders as dark gray, reflecting its intermediate attenuation.

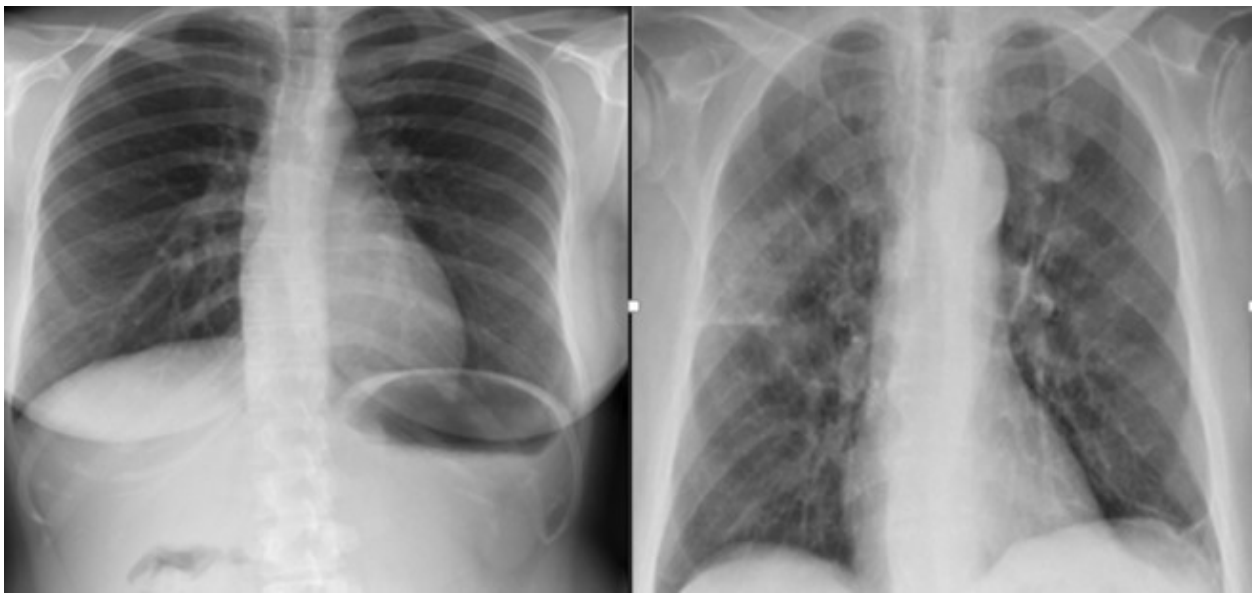

Figure 5: Comparison of a healthy lung (left) versus a pathological lung (right).

- **Soft tissues (e.g., heart, blood vessels):** Present as light gray, indicating higher density than fat but lower than bone.
- **Bone (e.g., ribs, vertebrae):** Shows up as whitish regions, since bone absorbs more X-ray photons.
- **Metallic implants (e.g., pacemakers, surgical clips):** Appear as very bright white due to extreme radiopacity.

These density gradients facilitate visual inspection: ribs and other osseous structures stand out in lighter shades, while aerated lung parenchyma is represented by near-black intensities. In a normal lung X-ray, the predominance of dark pixels corresponds to air-filled alveoli. Conversely, pathological conditions—such as pneumonia or pulmonary edema—introduce fluid or consolidation, pushing pixel intensities toward mid-gray as water density replaces air in affected regions. This results in blurred lung boundaries and loss of the expected sharp contrast between lung fields and surrounding tissues.

## 6 Experiments

All experiments utilize DenseNet201, selected for its effective feature propagation and its reported COVID-19 detection accuracy exceeding 95% [20]. Section 6.1 presents the experimental settings of the proposed methods. Section 6.1.1 establishes the baseline by classifying full chest X-ray images. Sections 6.1.2 and 6.1.3 then introduce occlusion-based classifiers trained on 4×4 and 3×3 patch grids, respectively. Sections 6.1.4 and 6.1.5 describes the Variational and Perceptual Autoencoder models trained exclusively on healthy lungs to produce anomaly maps for explanation. Section 6.2 presents global performance results, comparing all methods quantitatively using previously trained models of patches and VAE. Finally, Section 6.3 provides qualitative case studies, illustrating the spatial explanations generated by each approach.

In the global evaluation (Section 6.2), we also add a GradCAM-based baseline [25] using our DenseNet201 backbone. At inference, DenseNet201 generates a pixel-wise importance map via a gradient-based attribution technique, where higher attribution values indicate pixels belonging to the region of interest. To incorporate this into our 4×4 grid framework, we sum the attribution scores within each cell and select the cell with the highest total. This discretization ensures that GradCAM's often-irregular saliency patterns are mapped onto the same patch-based structure used by our other methods, allowing for a fair, grid-aligned comparison. In contrast, for the PAE-based variant we *do not* enforce a grid discretization, since the goal is to evaluate occlusion through a direct pixel-wise mask derived from reconstruction error (i.e., masking the entire detected region rather than a single selected cell).

For the global evaluation (Section 6.2), we employ four metrics to assess both classification performance and explanation fidelity: the probability difference ($\Delta p$), the sum of predicted probabilities, binary cross-entropy (BCE), and the sum of squared errors (SSE). In case of BCE, we use the typical definition where the predicted model based on occlusion is compared to real label instead to output of model using full images.

Given $N$ as the number of test images, $y^{(i)}_{\text{model\_pred}}$ the prediction of model based on occlusion, $y^{(i)}_{\text{cnn\_pred}}$ the prediction of model considering full images, and $y_{\text{true}}$ the true label of images, these metrics are defined as follows:

$$\Delta p \;=\; \sum_{i=1}^{N} \big(y^{(i)}_{\text{model\_pred}} \;-\; y^{(i)}_{\text{cnn\_pred}}\big)$$

$$\text{Sum of probabilities} \;=\; \sum_{i=1}^{N} y^{(i)}_{\text{model\_pred}}$$

$$\text{SSE} \;=\; \sum_{i=1}^{N} \big(y^{(i)}_{\text{cnn\_pred}} - y^{(i)}_{\text{model\_pred}}\big)^2$$

$$\text{BCE} = \text{Binary Cross-Entropy}\big(y_{\text{true}},\, y_{\text{model\_pred}}\big)$$

In our experiments, we encode the healthy class as 1 and the COVID-19 class as 0. For COVID-positive cases (class 0), an effective occlusion will produce a large probability difference ($\Delta p$) by substantially lowering the model's confidence in class 0. Similarly, the total sum of predicted COVID probabilities will increase, and the sum of squared errors (SSE) between original and occluded predictions also will increase, both indicating heightened model confusion. We add binary cross-entropy (BCE), which will increase when the model's output shifts away from the true label (0), reflecting the reduced certainty in its original prediction. In the case of COVID-negative instances (class 1), we hypothesize that occlusion will not significantly affect the model's performance; however, these cases are still evaluated using the global metrics. The confusion matrices characterize the classifiers used in the explanation pipeline, while the main comparison focuses on occlusion-based explanation metrics. A broader diagnostic evaluation with AUC-ROC, F1-score, and confidence intervals is left as future work.

### 6.1 Experimental settings

#### *6.1.1 COVID Classification on Full Chest X-Rays*

All chest radiographs were first preprocessed via histogram equalization to improve contrast and standardize intensity values across images. These full-frame X-rays were then input to a DenseNet201 backbone that had been fine-tuned on a stratified cohort of 3,034 COVID-positive and 7,925 healthy cases. The dataset was partitioned into 80% training, 10% validation, and 10% test splits, each preserving the original class proportions.

Training proceeded for up to 40 epochs with a batch size of 4, employing the Adam optimizer and a binary cross-entropy loss. We tracked validation accuracy and loss at each epoch, applying early stopping when no further validation improvement was observed over consecutive epochs.

On the held-out test set, the model demonstrates strong discriminative ability between COVID and non-COVID cases. Figure 6 presents the confusion matrix summarizing the classifier's performance on full-image inputs.

In summary, the DenseNet model demonstrates strong effectiveness in COVID classification; however, it does not explicitly reveal which regions it attends to. Below, we introduce the models proposed to localize these diagnostically relevant areas.

#### *6.1.2 COVID Classification Model considering 4×4 Patches*

To identify the most diagnostically relevant regions in chest X-rays, we conducted a patch-based analysis using a DenseNet201 backbone [20] considering procedure given in Section 4.1 (Steps 1 and 2). Each input image was divided into a $4 \times 4$ grid of non-overlapping patches, where each patch measured $56 \times 56$ pixels. These 16 patches were independently processed by a DenseNet201 classifier pretrained and fine-tuned to detect COVID-19. The network output for each patch was a scalar score proportional to the likelihood of COVID-19 presence within that patch; higher scores indicate greater relevance to the overall diagnosis.

For this experiment, the extracted patches comprised 3,616 COVID-positive patches and 10,192 healthy lung patches. We as splits the partitions of training (80%), validation (10%), and test (10%) sets, maintaining class balance in each subset. During training, we used a batch size of 12, the Adam optimizer, and a binary cross-entropy loss function. We trained for up to 40 epochs, monitoring both validation accuracy and validation loss, and applied early stopping if no improvement was seen over several consecutive epochs.

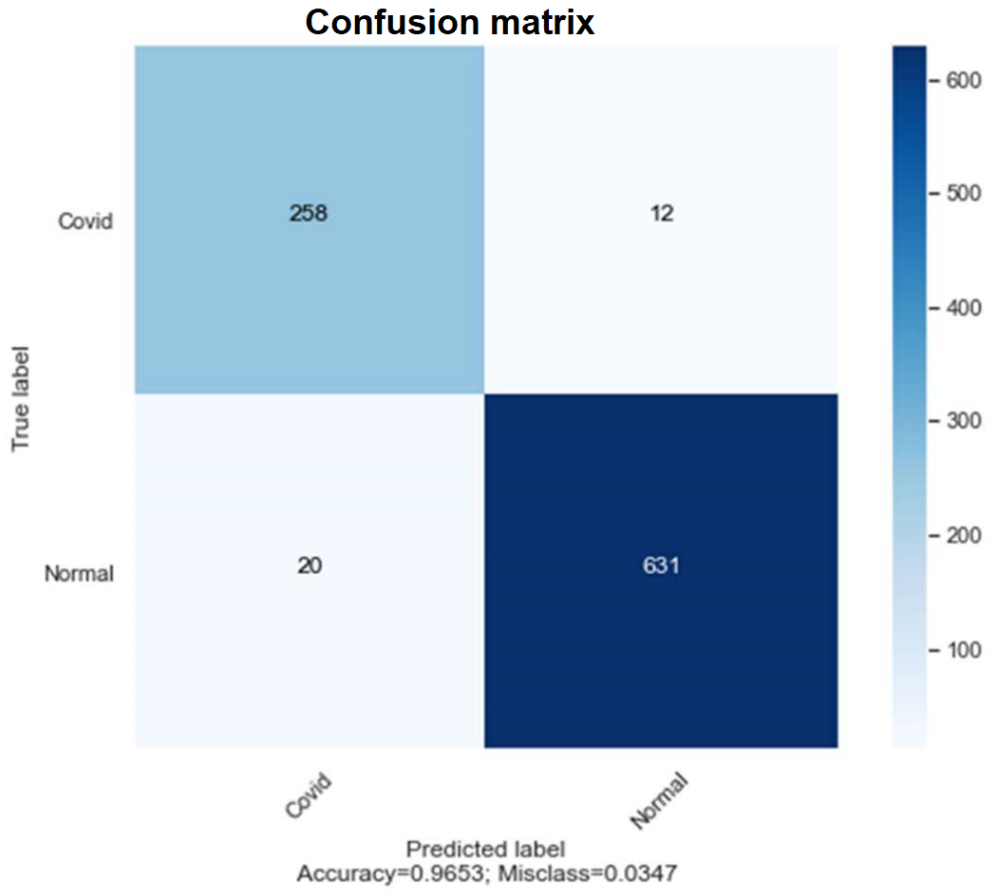


Figure 6: Confusion matrix for DenseNet201 on full chest X-rays from the test set.

The model's performance on patch-level classification is summarized by the confusion matrix in Figure 7. Note that this task is more challenging than whole-image classification, since operating on patches provides the classifier with less contextual information; consequently, its performance is predictably lower.

We then applied this model following the procedure in Section 4.1, occluding the patch with the highest response; the resulting performance is reported in Section 6.2.

### *6.1.3 COVID Classification Model considering 3×3 Patches*

In this experiment, we consider again the procedure given in Section 4.1 (Steps 1 and 2) with different resolution. Each chest X-ray was divided into 9 independent images, referred to as patches, each measuring 75×75 pixels. These patches were generated by splitting the original image into a 3×3 grid. The resulting images were then used as input for a CNN model based on DenseNet201, which was trained to detect the presence of COVID-19 in patients. The model was configured to assign a value of 0 to patches from COVID-19 positive patients and a value of 1 to patches from patients with normal chest images. The training was carried out using a dataset composed of 13,808 images, evenly distributed between both groups.

In this case, the dataset used in the experiment contained 3,616 patches from COVID-affected lungs and 10,192 patches from healthy lungs. The training, validation, and test sets were split in a stratified manner with 80%, 10%, and 10% of the data, respectively.

During the training phase, data augmentation techniques were applied to increase the variability of the input data and enhance the performance of the learning process. The model was trained for 30 epochs using a batch size of 30, the Adam optimizer, and the Binary Cross-Entropy loss function. The outcome of the training is presented in the confusion matrix given in Figure 8.

We observe a slight improvement relative to the previous experiment, which is reasonable given the greater amount of available visual information. As previous experiment, we then applied this classifier following the procedure in Section 4.1, occluding the patch with the highest response; the resulting performance is also reported in Section 6.2.

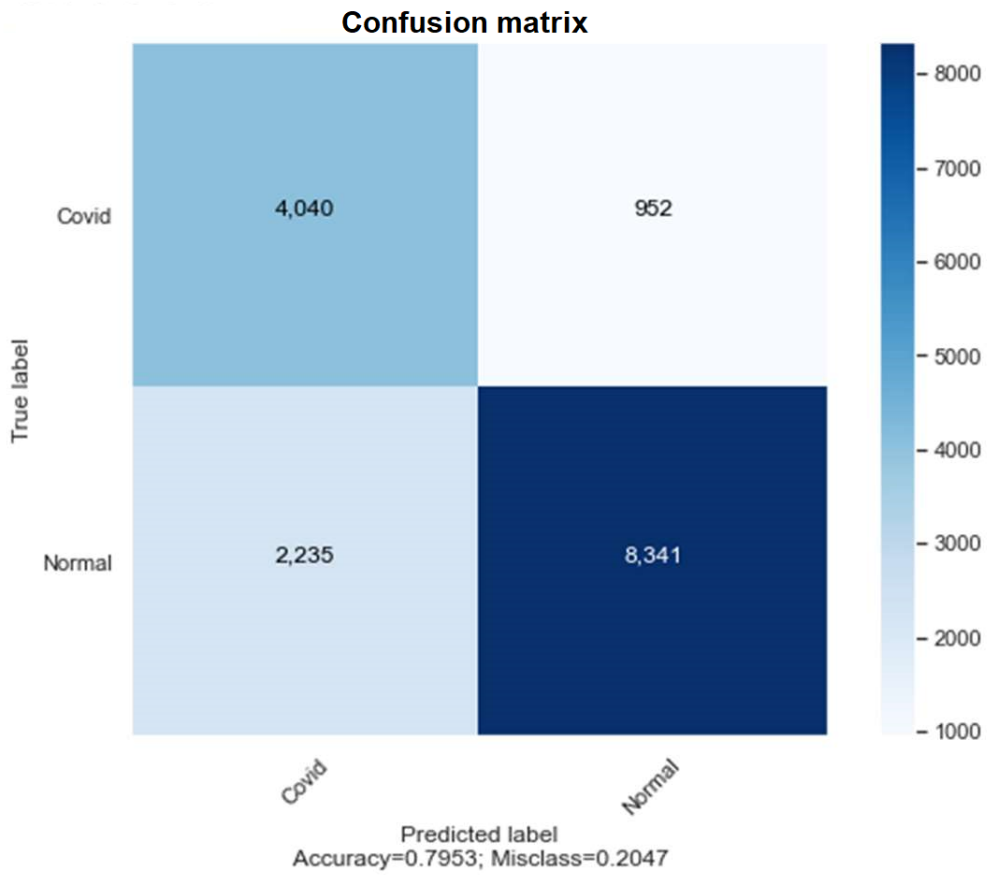


Figure 7: Confusion matrix for 4x4-patch-based DenseNet201 classification on the test set.

#### *6.1.4 Variational Autoencoder Training*

In this experiment, we consider again the procedure given in Section 4.2 (Step 2). After preprocessing the images through normalization, histogram equalization, and splitting into training, validation, and test sets, the dataset was distributed as follows where a total of 13,550 images were used for training, distributed accordingly. The $\beta$ value was kept fixed during the experiments (0.1). Since the goal of this work is to compare explanation strategies rather than optimize the VAE hyperparameter space, no $\beta$-ablation study is reported.

During training, data augmentation techniques were applied to increase the variability of the input and improve the performance of the learning algorithm. The model was trained over 1,000 epochs with a batch size of 24, using the Adam optimizer and the Binary Cross-Entropy loss function. The model converges in training and validation sets.

As previous experiment, we then applied this model following the rest of steps of procedure in Section 4.2, occluding the patch with the highest response; the resulting performance is also reported in Section 6.2.

#### *6.1.5 Perceptual Autoencoder Training*

In this experiment, we consider again the procedure given in Section 4.3 (Step 2). After preprocessing the images through normalization, histogram equalization, and splitting into training, validation, and test sets, the dataset was distributed as follows, where a total of 13,550 images were used for training (distributed accordingly).

During training, data augmentation techniques were applied to increase the variability of the input and improve the performance of the learning algorithm. The PAE backbone is a convolutional encoder–decoder with a strict bottleneck and no long-range skip connections (to prevent copying anomalies through shortcuts). The decoder produces outputs via a `tanh` activation, hence reconstructions lie in $[-1, 1]$ and the input normalization is kept consistent with this range. The model was trained over 1,000 epochs with a batch size of 24, using the Adam optimizer and a reconstruction objective that combines a pixel-wise term

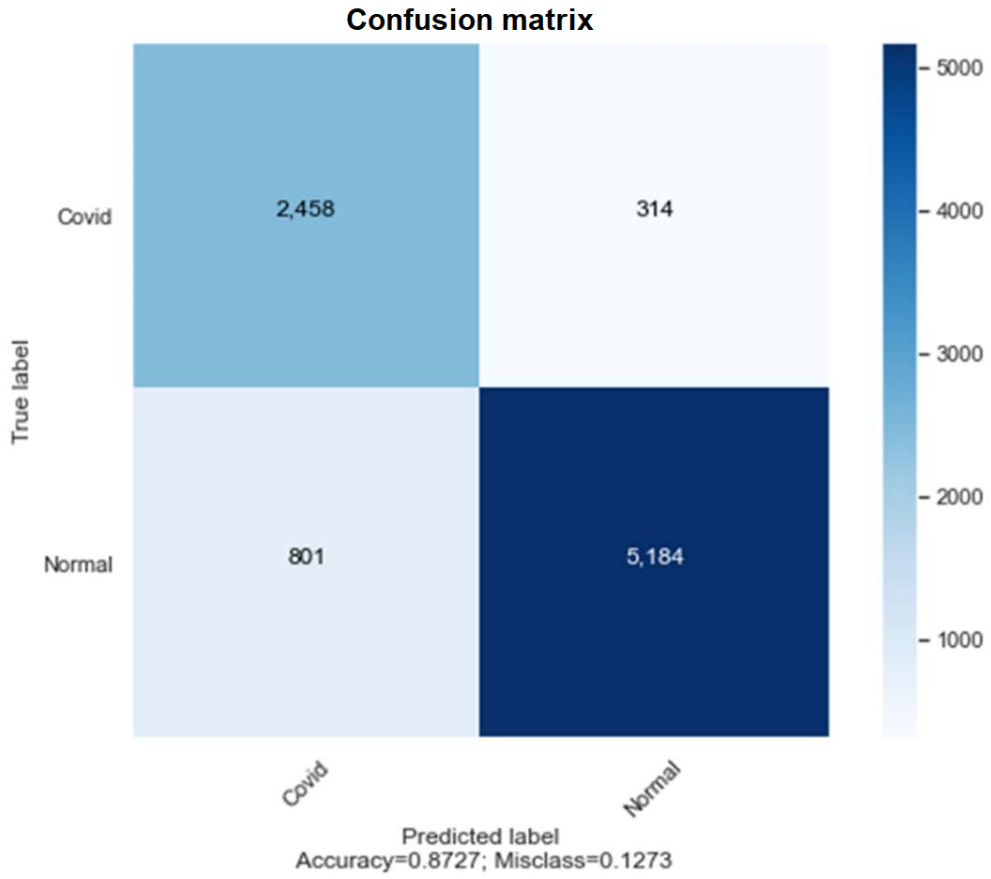


Figure 8: Confusion matrix for DenseNet201 using 3x3 patch-based classification.

and a perceptual similarity term:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_1(x, \hat{x}) + \lambda \, \mathcal{L}_{\text{LPIPS}}(x, \hat{x}).$$

The model appears to converge in training and validation sets. As in the previous experiment, we then applied this model following the remaining steps of the procedure in Section 4.3; however, in the PAE variant the occlusion is driven by a pixel-wise mask derived from the reconstruction error (adaptive thresholding), rather than selecting a single grid patch. The resulting performance is also reported in Section 6.2.

### 6.2 Global quantitative results

In this section, we present the results of applying the baseline GradCAM method using the DenseNet model, the proposed patch-based method with 3×3 and 4×4 configurations, and the VAE-based method. For the VAE and PAE approaches, in Step 2 of Sections 4.2 and 4.3, we use the DenseNet model obtained in Section 6.1.1. Below, we report the quantitative results of the four explanation techniques evaluated at the global level.

Table 1: Global Results of the Evaluated Visual Explainability Techniques

| **Model / Metric** | Δ Prob (↑) | Sum Prob (↑) | BCE (↑) | SSE (↑) |
|---|---|---|---|---|
| GradCAM | 6.77 | 10.46 | 0.178 | 2.66 |
| Patch 4×4 | 6.14 | 9.83 | 0.182 | 2.59 |
| Patch 3×3 | 6.02 | 9.71 | 0.181 | 2.13 |
| VAE | 7.27 | 10.96 | 0.177 | 2.87 |
| PAE | **10.57** | **12.04** | **0.454** | **7.88** |

Table 1 presents the comparative evaluation of the explainability techniques on the chest X-ray test dataset; the arrow orientation in the caption denotes the preferred direction for each metric. Notably, the

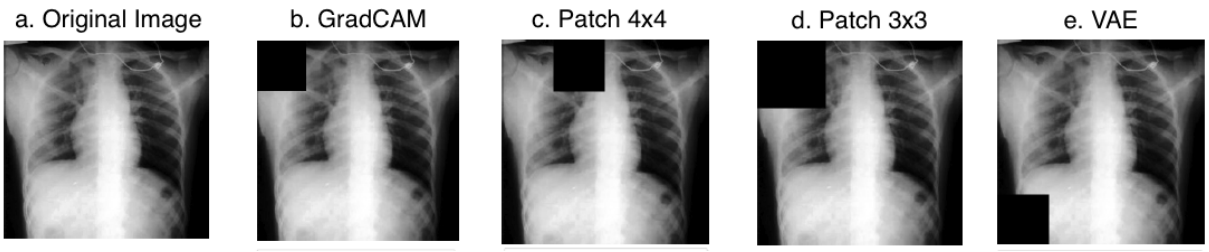


Figure 9: Visual results for evaluated models on confirmed COVID-19 case (Case 1).


Figure 10: Visual results for evaluated models on confirmed healthy lung case (Case 2).

**PAE-based** method achieves the best overall results, attaining the largest ΔProb (10.57) and the highest Sum Prob (12.04), while also producing the highest BCE and SSE (0.454, 7.88), consistent with the strongest prediction shift after occluding the highlighted evidence. The **VAE-based** method is the second-best overall, with the next highest ΔProb (7.27) and Sum Prob (10.96). Overall, the results indicate that autoencoder-based explainability (VAE/PAE) outperforms GradCAM and patch-only baselines, suggesting more faithful localization of diagnostically relevant regions. In addition, the mask-based occlusion used by PAE yields stronger metric gains than patch-based occlusion, although it provides less discrete, grid-localized attribution.

Although the VAE-based method obtains higher ΔProb and Sum Prob than GradCAM, this difference is relatively small and should be interpreted cautiously. Since repeated independent runs were not part of the experimental protocol, we do not claim statistical superiority for small numerical differences. Instead, these results are interpreted as descriptive evidence that reconstruction-based explanations are competitive with activation-based and patch-based alternatives.

### 6.3 Qualitative Analysis

Figures 9 and 10 illustrate two representative scenarios: Case 1 (confirmed COVID-19) and Case 2 (healthy lung), respectively. In Case 2, where bilateral basal opacities are evident, both GradCAM and patch-based occlusion highlight broad areas, but only the VAE-derived anomaly map accurately delineates the lesions with high spatial precision. Occluding the VAE-selected patch in this case produces the largest probability difference ($\Delta p = 25.7\%$), compared to $\Delta p \approx 15.3\%$ for GradCAM and $\approx 6.4\%$ for the 3×3 patch method, underscoring the VAE pipeline's superior localization and quantitative clarity. Case 2 is shown primarily for completeness: all methods generate minimal or diffuse signals—GradCAM and patch occlusion yield scattered highlights, while the VAE reconstruction error remains uniformly low—resulting in negligible change in model confidence. This confirms that even in non-pathological cases, the methods behave consistently, although our framework is primarily designed to explain positive findings.

Additionally, Figures 11 and 12 provide complementary qualitative evidence for the same masking principle using the proposed PAE. Unlike the patch-based variants in the previous analysis, the PAE approach occludes evidence without selecting fixed patches, for such reason we separate its qualitative analysis. In the Normal controls (top rows), the PAE reconstruction remains close to the input, yielding only low-amplitude, texture-like residuals and an error mask that produces no change in prediction ($P(\text{Covid}) = 0.0\%$ before and after occlusion). In contrast, for the COVID-positive examples (bottom rows), the PAE reconstruction suppresses disease-related patterns, so the residual concentrates over abnormal pulmonary regions; occluding these high-error areas leads to a sharp confidence drop (e.g., $99.6\% \rightarrow 11.8\%$ and $99.5\% \rightarrow 0.0\%$), indicating that the PAE-identified regions capture most of the discriminative evidence.

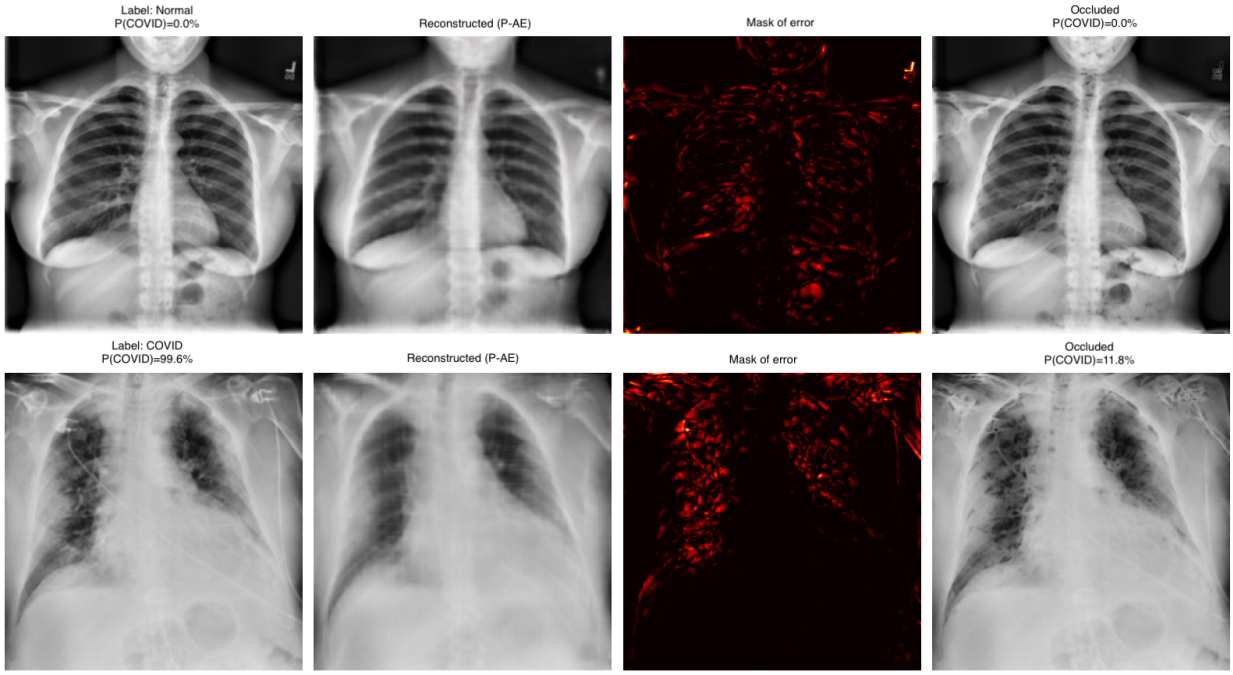


Figure 11: Visual results for evaluated models on confirmed COVID-19 case (Case 3).

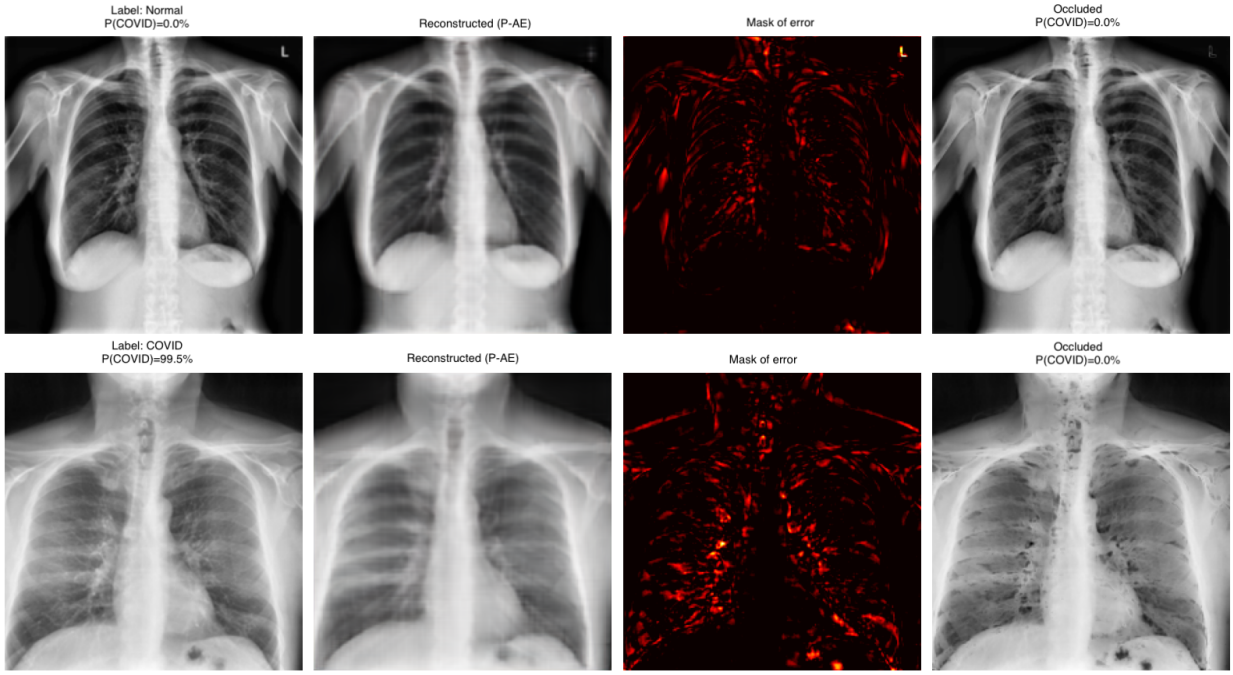


Figure 12: Visual results for evaluated models on confirmed healthy lung case (Case 4).

## 7 Conclusions

This work proposes an explainability framework for COVID-19 classification in chest X-rays that couples a DenseNet backbone with reconstruction-driven anomaly discovery and perturbation-based attribution. By training generative models exclusively on Normal images, reconstruction residuals provide a principled signal to localize deviations from healthy appearance, and the subsequent occlusion step offers a quantitative sanity check by measuring how strongly the classifier depends on the highlighted evidence.

Across the global evaluation, autoencoder-based explainability consistently outperforms GradCAM and patch-only occlusion. In particular, the VAE-based approach improves over both grid baselines, while the

PAE-based variant achieves the strongest overall effect: its perceptual reconstruction produces more contrastive residuals, and the resulting pixel-wise mask yields the largest prediction shifts after occlusion. These results indicate that reconstruction-error maps provide more faithful localization of diagnostically relevant regions than activation-only visualizations or fixed-grid masking, especially when the objective is to identify evidence whose removal substantially reduces COVID confidence.

Qualitatively, the VAE-derived anomaly maps better delineate lesion-like regions in positive cases compared to GradCAM and patch occlusion, while remaining near-uniform in healthy controls, producing negligible confidence changes as expected. Complementarily, the PAE provides an alternative occlusion mechanism that does not rely on selecting fixed patches: masking the regions identified by the reconstruction error can drive large drops in confidence in COVID-positive examples, supporting the interpretation that these regions carry most of the discriminative evidence.

For future work, we plan to (i) further study the trade-off between reconstruction fidelity and anomaly contrast by tuning VAE regularization (e.g., the $\beta$ parameter) and the PAE perceptual weight, (ii) analyze the stability of the pixel-wise masking strategy (e.g., thresholding rules and morphological refinement) to reduce sensitivity to acquisition artifacts, and (iii) investigate representation overlap between COVID and Normal images in latent space, aiming to improve separation and robustness under dataset shifts.

## Acknowledgment

B. Peralta appreciates the support of the National Center for Artificial Intelligence CENIA FB210017, Basal ANID and Chilean National Agency for Research and Development (ANID), Fondecyt grant ID 1241882.

## References

[1] PLOS Biology, "A positive feedback synapse from retinal horizontal cells to cone photoreceptors," *PLOS Biology*, May 03 2011.

[2] Organización Mundial de la Salud, "Brote de enfermedad por coronavirus (covid-19)," https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019, n.d.

[3] Gobierno de Chile, "Cifras oficiales del plan paso a paso," n.d. [Online]. Available: https://www.gob.cl/pasoapaso/cifrasoficiales/

[4] National Library of Medicine, "Diagnóstico radiológico del paciente con covid-19," *National Library of Medicine (NCBI)*, Nov 24 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7685043/

[5] European Journal of Radiology, "Artificial intelligence for imaging-based covid-19 detection: Systematic review comparing added value of ai versus human readers," *European Journal of Radiology*, Nov 16 2021. [Online]. Available: https://doi.org/10.1016/j.ejrad.2021.110028

[6] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[7] R. Bayuk, J. Manquel, O. Nicolis, L. Caro, and B. Peralta, "Explainable covid-19 classification via variational autoencoder-guided patch occlusion," in *2025 LI Latin American Computer Conference (CLEI)*. IEEE, 2025, pp. 1–10.

[8] I. Allaouzi and M. Ahmed, "A novel approach for multi-label chest x-ray classification of common thorax diseases," *IEEE Access*, vol. PP, pp. 1–1, 05 2019.

[9] M. Hanif, M. Bilal, A. Alsaggaf, and U. Al-Saggaf, "Enhancing multi-label chest x-ray classification using an improved ranking loss," *Bioengineering*, vol. 12, no. 6, p. 593, 2025. [Online]. Available: https://doi.org/10.3390/bioengineering12060593

[10] X. Zhao and X. Wang, "Multi-label chest x-ray image classification based on long-range dependencies capture and label relationships learning," *Biomedical Signal Processing and Control*, vol. 100, p. 107018, 2025.

[11] M. Chetoui, M. Akhloufi, B. Yousefi, and E. Bouattane, "Explainable covid-19 detection on chest x-rays using an end-to-end deep convolutional neural network architecture," *Big Data and Cognitive Computing*, vol. 5, no. 4, p. 73, 2021.

[12] J. Rocha, S. Pereira, P. Sousa *et al.*, "Clare-xr: Explainable regression-based classification of chest radiographs with label embeddings," *Scientific Reports*, vol. 14, p. 31024, 2024.

[13] R. F. Mansour, J. Escorcia-Gutierrez, M. Gamarra, D. Gupta, O. Castillo, and S. Kumar, "Unsupervised deep learning based variational autoencoder model for covid-19 diagnosis and classification," *Pattern Recognition Letters*, vol. 151, pp. 267–274, 2021.

[14] D. Addo, S. Zhou, J. Jackson, G. Nneji, H. Monday, K. Sarpong, R. Patamia, F. Ekong, and C. Owusu-Agyei, "Evae-net: An ensemble variational autoencoder deep learning network for covid-19 classification based on chest x-ray images," *Diagnostics*, vol. 12, no. 11, p. 2569, 2022.

[15] Q. Zhou, S. Wang, X. Zhang, and Y. Zhang, "Wvale: Weak variational autoencoder for localisation and enhancement of covid-19 lung infections," *Computer Methods and Programs in Biomedicine*, vol. 221, p. 106883, June 2022.

[16] N. Jahan and M. A. Hasan, "Autoencoder-based unsupervised anomaly detection for covid-19 screening on chest x-ray images," in *2022 19th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, 11 2022, pp. 1–6.

[17] H. Guan, P.-T. Yap, A. Bozoki, and M. Liu, "Federated learning for medical image analysis: A survey," *Pattern recognition*, vol. 151, p. 110424, 2024.

[18] X. Chen, X. Wang, K. Zhang, K. Fung, T. Thai, K. Moore, R. Mannel, H. Liu, B. Zheng, and Y. Qiu, "Recent advances and clinical applications of deep learning in medical image analysis," *Medical Image Analysis*, vol. 79, p. 102444, July 2022.

[19] J. Martínez, O. Nicolis, L. Caro, and B. Peralta, "A proposal for the deep unsupervised identification of relevant areas in x-rays for covid detection," in *2021 IEEE Chilean Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON)*. IEEE, 2021, pp. 1–6.

[20] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[21] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations*, 2014.

[22] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[23] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[24] T. Rahman, S. C. H. Hossain, M. Shishir *et al.*, "Covid-19 radiography database," https://www.kaggle.com/tawsifurrahman/covid19-radiography-database, 2020, accessed: 2025-06-05.

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.