VNIVERSITAT DE VALÈNCIA

Facultat de Física

Master's Degree in Medical Physics

VNIVERSITAT DE VALÈNCIA

Master's Thesis

Application of Deep Learning Algorithms for Medical Image Classification

Alejandro Ruiz Martínez

Supervisor: Joan Vila Francés

València, September 23, 2025

D/Dña **Joan Vila Francés**, Profesor Titular del Dpto. de **Ing. Electrónica** de la Universitat de **València**

CERTIFICA/N:

Que la presente memoria, titulada "**Aplicación de algoritmos de Deep Learning para la clasificación de imágenes médicas**", corresponde al trabajo realizado bajo su dirección por D/Dña. **Alejandro Ruiz Martínez**, para su presentación como Trabajo Fin de Máster en el Máster Universitario en Física Médica de la Universitat de València.

Y para que conste firma/n el presente certificado en Valencia, a 2 de septiembre de 2025.

Fdo. Joan Vila Francés

Contents

Glossary of Acronyms

AI – Artificial Intelligence

ML – Machine Learning

DL – Deep Learning

CNN – Convolutional Neural Network

BCE – Binary Cross-Entropy

PA – Posteroanterior

API – Application Programming Interface

TPR/FPR – True Positive Rate / False Positive Rate

CAM / Grad-CAM – Class Activation Map / Gradient-weighted Class Activation Mapping

TL – Transfer Learning

Abstract

This thesis addresses multi-label classification of PA-projection chest radiographs from PadChest using convolutional neural networks. We conduct a controlled evaluation of architectural and training factors for the classifiers. The primary metrics for model comparison are macro-averaged AUC-ROC (area under the ROC curve) and PR-AUC (area under the Precision–Recall curve); additionally, we report calibration (Brier score), clinically motivated operating points (Spec@Sens=95% and Sens@Spec=95%), and interpretability via Grad-CAM maps.

Results indicate that performance is driven more by the training procedure than by the choice of backbone family. Compared with prior work on PadChest, the final model shows improved performance on structural findings (e.g., *Scoliosis*, *Vertebral degenerative changes*) while being more conservative on diffuse findings (e.g., *Pleural effusion*). Overall, we establish a reproducible, clinically oriented baseline that jointly reports discrimination, calibration, variability, and interpretability.

1 Introduction

Artificial Intelligence (AI) has evolved from an academic field into a cross-cutting technology that is transforming how we produce, diagnose, and make decisions. In this context, its adoption in healthcare—especially in medical imaging—creates opportunities to improve efficiency, reproducibility, and access to high-quality diagnostics.

AI can be understood as the effort to automate intellectual tasks performed by humans. In contrast to symbolic AI—based on explicit rules and knowledge bases—data-driven approaches learn directly from examples, enabling us to tackle problems with strong perceptual and interpretive components.

Within AI, machine learning (ML) comprises algorithms that learn the rules mapping data to solutions. Specifically, these algorithms learn the parameters of functions that relate inputs to desired outputs from labeled examples [1]. The goal is to discover representations that facilitate solving the target task [2]. Such representations are useful transformations of the data that reduce problem complexity and are fitted through training on real-world datasets. Unlike classical programming—where the programmer explicitly codes the rules that transform the data—ML systems learn those rules automatically from the data.

Deep learning (DL) is a subfield of ML that uses deeper models capable of extracting increasingly hierarchical representations useful for complex tasks [3]. In this work, DL is the methodological backbone chosen for its ability to learn directly from large collections of medical images.

Medical imaging is a particularly suitable domain for DL due to the growing availability of datasets, the visual nature of findings, and the need for consistent decisions. In diagnostic imaging, DL models are commonly applied to classification, detection, segmentation, registration, and generation across areas such as the chest, brain, retina, and digital pathology [4].

Convolutional neural networks (CNNs) are the reference architecture in DL for computer vision. CNNs apply learned filters that capture relevant spatial patterns [5], enabling recognition of the presence or absence of findings. This architecture has proven effective in large-scale visual recognition [6, 7] and forms the basis of the models evaluated in this work.

This Master's Thesis addresses multi-label classification of chest radiographs: a setting in which each study may present multiple findings simultaneously and, therefore, each label is modeled as an independent binary classifier. This approach mirrors clinical practice, where a single report describes multiple findings.

Mathematical Framework

ML and DL algorithms are called neural networks. A neural network is composed of interconnected layers of neurons (differentiable functions). These neurons have trainable parameters—weights and biases [8]. The first layer receives the data and the last produces an output appropriate to the task. In this work we address binary multi-label classification of radiographic images using convolutional neural network (CNN) architectures.

In this kind of network, the initial input is an image represented as a three-dimensional array $X^{(0)} \in \mathbb{R}^{H_0 \times W_0 \times C_0}$, where H_0 and W_0 are height and width (in pixels), and C_0 is the number of channels (e.g., C_0=1 for grayscale or C_0=3 if channels are replicated). At layer l, C_l filters (or *kernels*)—typically square of size $K_l \times K_l$ and full depth C_{l-1}—are applied; each filter "sees" all C_{l-1} input channels simultaneously and produces a single output channel. We denote by S_l the *stride* (the horizontal/vertical step) and by P_l the *padding* (the number of pixels added around the image, typically zeros), which together control the spatial size of the output.

The spatial dimensions of the layer output are

$$H_l = \left[\frac{H_{l-1} + 2P_l - K_l}{S_l}\right] + 1, \qquad W_l = \left[\frac{W_{l-1} + 2P_l - K_l}{S_l}\right] + 1. \tag{1}$$

If $W^{(l,c)} \in \mathbb{R}^{K_l \times K_l \times C_{l-1}}$ are the weights of filter c and $b_c^{(l)} \in \mathbb{R}$ its bias, the output of each filter (channel c) is the discrete convolution between the input and the filter:

$$Z_{i,j,c}^{(l)} \;=\; \sum_{m=1}^{C_{l-1}} \sum_{u=1}^{K_l} \sum_{v=1}^{K_l} W_{u,v,m}^{(l,c)}\; X_{i\cdot S_l+u-P_l,\, j\cdot S_l+v-P_l,\, m}^{(l-1)} \;+\; b_c^{(l)}, \tag{2}$$

which defines the pre-activation feature map. Applying the activation function $\phi^{(l)}$ yields the activated feature map

$$A_{i,j,c}^{(l)} = \phi^{(l)}\Big(Z_{i,j,c}^{(l)}\Big), \tag{3}$$

and stacking the C_l channels produces the layer output $A^{(l)} \in \mathbb{R}^{H_l \times W_l \times C_l}$, which serves as the input to layer l+1. The number of trainable coefficients per layer is $K_l^2\, C_{l-1}\, C_l$ (weights) plus C_l (biases).

Activation functions introduce the nonlinearity required for the network to learn useful representations. The two options used in this work are the ReLU (*Rectified Linear Unit*)

$$\mathrm{ReLU}(z) = \max(0, z) = \begin{cases} 0, & z < 0, \\ z, & z \geq 0, \end{cases} \tag{4}$$

and the sigmoid

$$\sigma(z) = \frac{1}{1 + e^{-z}}, \tag{5}$$

with range $(0, 1)$, making it suitable to interpret outputs as probabilities. In binary

classification it is used in the output layer to map logits to per-label probabilities (as $p_{n,i} = \sigma(z_{n,i})$ in this work).

During training, images are fed to the model in fixed-size batches. After each layer processes all images in the batch, batch normalization (BN) [9] is applied:

$$\tilde{A}^{(l)}_{i,j,c,n} = \gamma_c \hat{A}_{i,j,c,n} + \beta_c, \tag{6}$$

where

$$\hat{A}_{i,j,c,n} = \frac{A^{(l)}_{i,j,c,n} - \mu_c}{\sqrt{\sigma_c^2 + \varepsilon}}, \tag{7}$$

and

$$\mu_c = \frac{1}{NH_lW_l} \sum_{i,j,n} A^{(l)}_{i,j,c,n}, \quad \sigma_c^2 = \frac{1}{NH_lW_l} \sum_{i,j,n} (A^{(l)}_{i,j,c,n} - \mu_c)^2. \tag{8}$$

The parameters γ_c and β_c, one per channel, are trainable like the rest of the weights and biases. Finally, *pooling* [10] is used to reduce the spatial dimensions of each layer's output. With a window $p \times p$ and stride S_p:

$$P^{(l)}_{i,j,c} = \max_{0 \le u,v < p} A^{(l)}_{i \cdot S_p + u,\, j \cdot S_p + v,\, c}. \tag{9}$$

This operation provides local translational invariance, reducing sensitivity to small input shifts and preventing unbounded growth in memory usage. The action of block l on the input $X^{(l)}$ is summarized as

$$X^{(\ell)} \xrightarrow{\text{Conv}+\phi+\text{BN}+\text{Pool}} X^{(\ell+1)}, \tag{10}$$

i.e., convolution, activation, batch normalization, and pooling.

The output of the final block, $X^{(L)} \in \mathbb{R}^{H_L \times W_L \times C_L}$, is reduced by global average pooling into a feature vector [11],

$$g_c = \frac{1}{H_L W_L} \sum_{i=1}^{H_L} \sum_{j=1}^{W_L} X^{(L)}_{i,j,c}, \quad \text{for } c = 1, \ldots, C_L. \tag{11}$$

The feature vector feeds a sequence of D fully connected layers. Each layer d consists of N_d neurons, each with weight vector $W^{(d,n)} \in \mathbb{R}^{N_{d-1}}$, bias $b^{(d,n)} \in \mathbb{R}$, and activation function $\phi^{(d)}$, producing the activations (logits before the final sigmoid, where applicable)

$$z^{(d)} = \phi^{(d)}(z^{(d-1)} W^{(d,n)} + b^{(d,n)}), \tag{12}$$

with $z^{(d-1)} = g_c$ for the first layer of the head.

For a multi-label classification problem, the last layer contains as many neurons as there

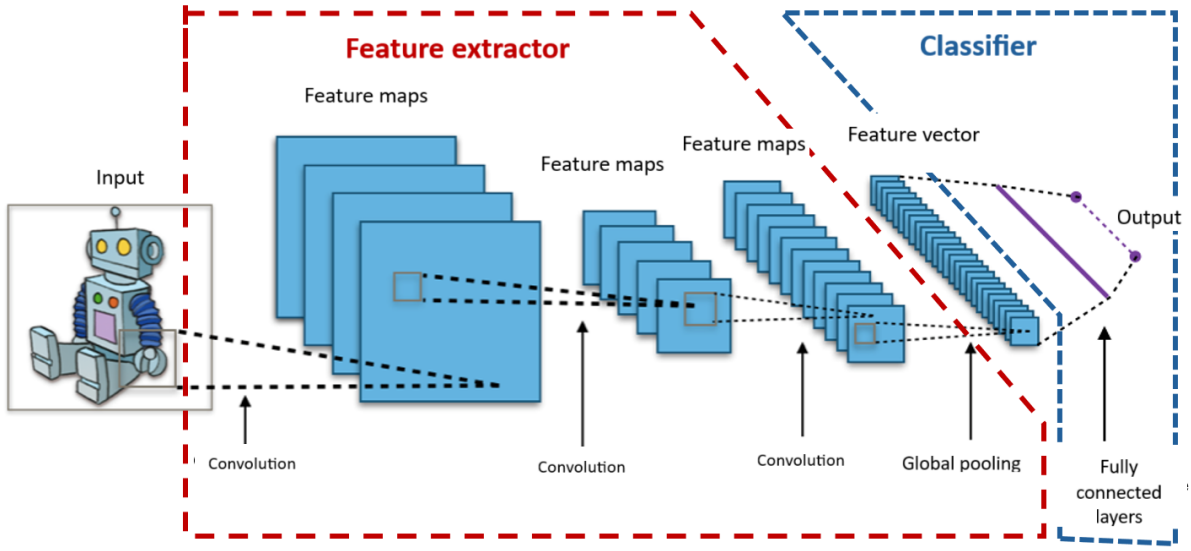


Figure 1: Typical CNN architecture. The label "Convolution" denotes the full action of a layer (convolution–activation–batch normalization–pooling). *Adapted from* Aphex34, *Typical cnn.png*, CC BY-SA 4.0.

are labels that can be assigned to each image, and outputs

$$p_i = \phi^{(D)}(h^{(D-1)}), \tag{13}$$

with $\phi^{(D)}() = \sigma()$ (the sigmoid function), so that $p_i \in [0, 1]$ is interpreted as the probability assigned to each class.

From these probabilities p_i and the ground-truth labels y_i for each class ($y_i = 0$ for the negative case and $y_i = 1$ for the positive case of class i), we define the loss function $\ell(y, p)$, which measures prediction quality on each batch. The weights and biases are updated in the direction of the negative gradient of $\ell(y, p)$ according to

$$w \leftarrow w \; - \; \eta \, \frac{1}{B} \sum_{n=1}^{B} \nabla_w \, \ell\Big(y^{(n)}, p^{(n)}\Big) \,, \tag{14}$$

in a process called backpropagation, with the objective of minimizing the loss [12]. The parameter η is the learning rate and determines the magnitude of the weight update. Proper tuning of this parameter enables convergence toward the (global) minimum of the loss.

The part of the algorithm that transforms the input images into g_c is called the feature extractor, and the set of fully connected layers acting on g_c is called the classifier head. A visual scheme of this type of algorithm is shown in Figure 1.

To obtain a positive or negative output for each label, we define a per-class threshold τ_i

such that the prediction becomes binary:

$$\widehat{y}_i = \begin{cases} 1 & \text{if } p_i \geq \tau_i, \\ 0 & \text{if } p_i < \tau_i. \end{cases} \tag{15}$$

Comparing $\widehat{y}_i$ with the ground-truth label y_i yields four cases: true positive (TP, $\widehat{y}_i$=1 and y_i=1), false positive (FP, $\widehat{y}_i$=1 and y_i=0), false negative (FN, $\widehat{y}_i$=0 and y_i=1), and true negative (TN, $\widehat{y}_i$=0 and y_i=0). These counts constitute the 2×2 confusion matrix for each label [13].

This classification procedure can engender a degree of distrust because it offers no explanation of the reasoning behind a decision. This is particularly critical in a domain as error-sensitive as medical imaging. Grad-CAM maps (gradient-weighted class activation mapping) attribute to each position of the input image its relative importance in the final decision. They are obtained from the class-specific gradients associated with the final convolutional layer [14]. When overlaid on the original image, they yield a heat map that highlights the regions most influential for the decision.

Evaluation Metrics

In this work we evaluate multi-label classifiers from per-label probabilities $p_i \in [0, 1]$. Our primary threshold-free metrics are the area under the Receiver Operating Characteristic curve (AUC-ROC) and the area under the Precision–Recall curve (PR-AUC). These metrics assess the model's ability to correctly rank cases without fixing a binary decision point. For clinical interpretability, for the best-performing model we also report sensitivity (recall) and specificity at operating points chosen by maximizing a given criterion, and we include a calibration measure via the Brier score [15].

At a given operating point (fixed threshold), using the four counts of the confusion matrix we define sensitivity (recall) and specificity as

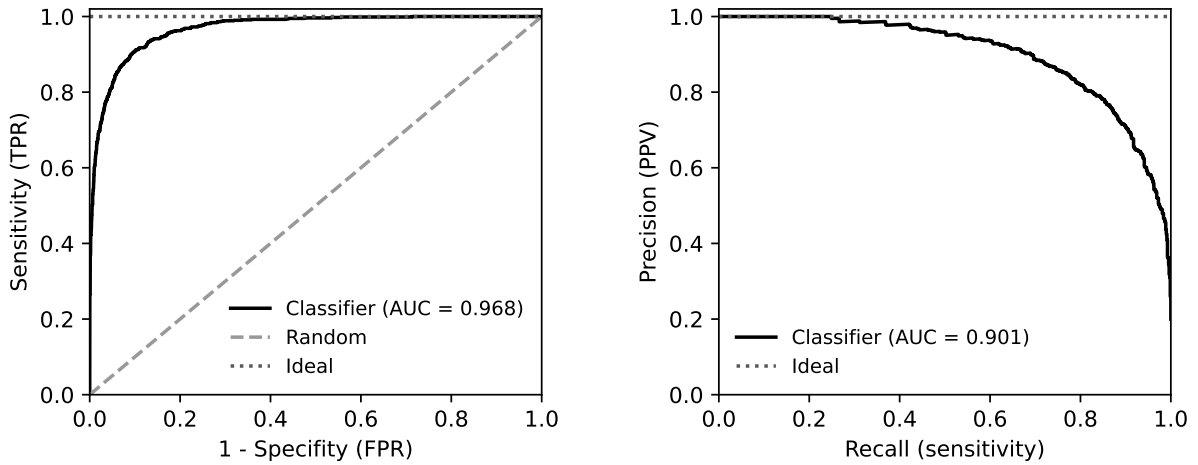
$$\text{Sens} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad \text{Spec} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{16}$$

and precision as

$$\text{Prec} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{17}$$

Intuitively, sensitivity measures how many true positives are detected, specificity how many true negatives are correctly rejected, and precision how many alarms are correct. All take values in $[0, 1]$; an ideal classifier would have Sens = Spec = Prec = 1, whereas low values indicate, respectively, missed positives, false alarms, or unreliable alerts. There is a trade-off as the threshold τ varies: increasing sensitivity typically decreases specificity, and with rare classes precision becomes more demanding.

The Receiver Operating Characteristic (ROC) curve plots, as the threshold varies, the



(a) ROC. X-axis: 1− Specificity (false positive rate, FPR). Y-axis: Sensitivity (true positive rate, TPR).

(b) PR. X-axis: Recall (Sensitivity). Y-axis: Precision (positive predictive value, PPV).

Figure 2: (a) ROC and (b) PR curves of an example classifier, together with the ideal case (dotted line) and the random case in (a) (dashed line).

pairs $\Big(1 - \text{Spec}, \text{Sens}\Big)$, as shown in Figure 2a. Its area (AUC-ROC) summarizes overall performance without relying on a specific threshold and can be interpreted as

$$\text{AUC-ROC} = \Pr\Big(p(x^+) > p(x^-)\Big), \tag{18}$$

i.e., the probability that the model score $p(\cdot)$ is higher for a positive x^+ than for a negative x^-. Intuitively, AUC-ROC measures the quality of ranking cases by likelihood of positivity. Its range is from 0.5 (random behavior) to 1 (perfect discrimination). The metric is relatively insensitive to prevalence, facilitating comparisons across labels or datasets with different prevalences [16].

The Precision–Recall (PR) curve relates, as the threshold varies, precision $\Big(\text{Prec}\Big)$ to recall/sensitivity $\Big(\text{Sens}\Big)$, as shown in Figure 2b. The area under this curve (PR-AUC) measures how "clean" the detections are while attempting to cover most positives: it simultaneously values few false alarms and high coverage. Its range is $[0, 1]$, and the baseline equals the label prevalence; with rare classes, maintaining high precision at high recall is especially difficult [17].

In a multi-label setting these metrics can be aggregated to provide a global view in two ways. Macro averaging gives equal weight to each finding. Micro averaging is obtained by summing TP, FP, FN, and TN across all labels and then applying the corresponding definition. Intuitively, macro reflects balanced performance across findings, whereas micro is dominated by the most frequent labels. For this reason, and for comparability with prior work, we report macro averages in this study.

Beyond discrimination, a good model should be calibrated: its predicted probabilities should correspond to observed frequencies. We quantify calibration with the Brier score,

defined as the mean squared error between predicted probabilities and ground-truth labels,

$$\text{Brier} = \frac{1}{N\,K} \sum_{n=1}^{N} \sum_{i=1}^{K} \Big(p_{n,i} - y_{n,i}\Big)^2, \tag{19}$$

where N is the number of images and K the number of labels. This metric takes values in $[0, 1]$; 0 indicates perfectly calibrated predictions and 1 the worst case. The Brier score penalizes both miscalibration and lack of sharpness (probabilities near 0.5), so lower values are preferable [15].

Statistical Comparison Between Models

To determine whether two models differ significantly on a given metric, we compare their metric values using Welch's t-test (two-sided, $\alpha = 0.05$) [18, 19]. Let k denote the number of replicates per model; given the sample mean $\bar{m}$ and sample standard deviation s for each model, the standard error of the difference in means and the test statistic are

$$SE = \sqrt{\frac{s_A^2}{k} + \frac{s_B^2}{k}}, \qquad t = \frac{\bar{m}_A - \bar{m}_B}{SE}, \tag{20}$$

and the effective degrees of freedom (Welch–Satterthwaite approximation) when $n_A = n_B = k$ are

$$df = (k-1)\,\frac{(s_A^2 + s_B^2)^2}{s_A^4 + s_B^4}. \tag{21}$$

We test $H_0 :\ \bar{m}_A = \bar{m}_B$ against $H_1 :\ \bar{m}_A \neq \bar{m}_B$. Let $t^* = t_{1-\alpha/2,\,df}$ be the 95% quantile of Student's t distribution; if $|t| > t^*$, we reject H_0 (there is a significant difference). Otherwise, we do not reject H_0.

Alternatively, the p-value can be computed as $p = 2 \cdot P(T > |t|)$, where T follows Student's t distribution with df degrees of freedom. If $p < 0.05$, we reject H_0. This is equivalent to the t^*-based rule, since $p < 0.05$ if and only if $|t| > t^*$.

2 Hypotheses and Objectives

This study pursues two primary objectives. First, it assesses the extent to which established conclusions from generic computer vision with deep CNNs transfer to the multi-label classification of findings in chest radiographs. Second, it explores specific architecture and training configurations that can improve state-of-the-art performance under realistic data and compute constraints, providing comparative evidence and practical criteria for eventual clinical use.

The first line of work compares two complementary axes: families of architectures ranging from very lightweight, simple networks to deeper, higher-capacity models, and training regimes that include both learning from scratch and fine-tuning after pretraining on natural images. We expect pretraining to yield richer representations and therefore better performance. The goal is to characterize the trade-offs among representational capacity, learning stability, and computational efficiency, and to identify reasonable configurations for the radiographic context.

To address the class imbalance characteristic of radiographic findings, we evaluate simple, reproducible strategies based on per-label weighting in the loss function and on sampling schemes. We examine which approach contributes most to performance on low-prevalence labels.

Additionally, to mitigate the imbalance between positive and negative cases and to promote generalization, we analyze data augmentation with geometric and intensity transformations compatible with chest radiography. Complementarily, we explore data-mixing strategies that combine pairs of images and their annotations in a controlled manner—either by interpolating intensities and targets or by inserting regions from one image into another—so that the resulting label reflects the mixing proportion.

Finally, we compare our results with the best published references on this same dataset. We further extend the analysis with information that is rarely reported in those references, in particular the assessment of probability calibration via the Brier score, together with operating points defined on validation for high-sensitivity and high-specificity scenarios that facilitate clinical implementation. We also assess clinical plausibility through an interpretability analysis using heat maps aimed at verifying consistency with radiological judgment, documenting transparency, and ruling out spurious dependencies.

3 Materials and Methods

Materials

The data come from PadChest by Bustos et al. [20], a chest radiograph dataset comprising over 160,000 studies with annotations of radiographic findings, diagnoses, and associated metadata. Approximately 27% of the images were manually annotated by radiologists, with the remainder labeled automatically via a natural language processing system that achieved a micro-F1 (harmonic mean of precision and recall) of 0.93 in its validation [20]. Consequently, label noise is present to some degree and must be considered in the experimental design and in the interpretation of results.

PadChest includes images acquired in different projections. A projection-wise breakdown shows that the posteroanterior (PA) view is the most common (91,728 cases), followed by the lateral view (L; 49,579 cases). Other projections—anteroposterior horizontal (14,346), standard anteroposterior (4,559), and rib (630)—are less frequent, and administrative categories such as *EXCLUDE* and *UNK* are rare (11 and 8 cases). To ensure dataset homogeneity, this study includes only PA-projection radiographs.

The label space comprises 174 unique categories with a marked frequency imbalance. Only five labels exceed 10,000 positive instances—*normal* (50,616), *COPD signs* (23,280), *cardiomegaly* (15,028), *unchanged* (14,349), and *aortic elongation* (10,824)—whereas at the other extreme some labels have very few samples—*pleural mass* (3), *esophagic dilatation* (3), *nephrostomy tube* (3), *lipomatosis* (1), and *breast mass* (1). It is also noteworthy that even the most prevalent radiographic finding (*cardiomegaly*) appears in fewer than 10% of studies, making the dataset substantially imbalanced toward the negative class across all labels considered.

Experiments were conducted in Python using TensorFlow [21] and its high-level Keras API [22] for model specification, training, and evaluation. This ecosystem provides the necessary components to reproduce the models, manage the training loop, and export per-label probabilities required by the evaluation protocol.

For transfer learning experiments, we used convolutional networks pretrained on *ImageNet* [23], a natural-image dataset for single-label multiclass classification. In particular, we considered three representative families spanning lightweight, low-complexity models to deeper, higher-capacity architectures. Among them, we evaluated configurations based on MobileNetV2 [24], ResNetV2 [25], and EfficientNet-B0 [26]. At a high level, MobileNet-type models employ inverted residual blocks with extensive use of depthwise separable convolutions to reduce computational cost; ResNet-type models use shortcut connections and preactivation to facilitate optimization in deeper networks; and EfficientNet starts from a base configuration (EfficientNet-B0) found via architecture search and built with MBConv blocks and squeeze-and-excitation attention.

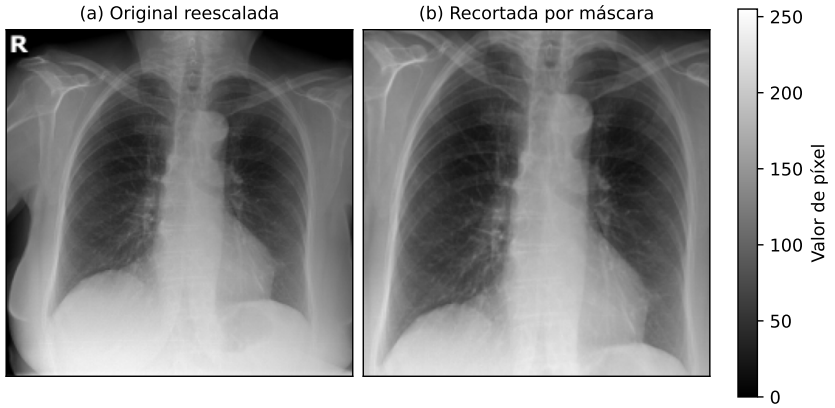


Figure 3: Example PA radiograph before (a) and after (b) lung-area cropping.

Methods

The experimental pipeline was designed to compare, in a controlled and reproducible manner, architecture and training configurations for multi-label classification of chest radiographs. The protocol comprises label preprocessing and encoding, patient-level splitting, definition of model families (shallow networks and deep transfer models), controlled activation of experimental factors aligned with the hypotheses, model training with the corresponding factors, and evaluation using threshold-free metrics and validation-defined operating points.

Image preprocessing consisted of lung-area cropping, conversion to grayscale, geometric rescaling to a fixed resolution of 224×224 pixels, and intensity normalization. For pretrained models, the single channel was replicated to obtain three-channel inputs and standard *ImageNet* normalization was applied to the rescaled values; for models trained from scratch, intensities in $[0, 1]$ were used after linear scaling.

Specifically, for lung cropping, PA-projection images were first rescaled to 512×512 to obtain a mask of both lungs using TorchXRayVision (pretrained PSPNet) [27, 28]. Images were cropped to the smallest square region enclosing all positive mask values, with an added margin of 3% of the size, provided the resulting size exceeded 50,000 pixels to avoid excessive cropping due to mask errors. An example of this transformation is shown in Figure 3.

Each study's labels were represented by a binary multi-label vector $y \in \{0, 1\}^K$, with K the number of findings considered. Each component y_i indicates presence (1) or absence (0) of finding i in the associated report. The absence of an explicit label was treated as negative.

To prevent information leakage, the split was performed at the *patient* level into training, validation, and test subsets. All decisions regarding threshold selection, early stopping, and hyperparameter tuning were made exclusively on the validation set; the test set was used only for post-training comparisons across models. All metrics reported in the Results section refer to performance on the test set.

We implemented a family of shallow networks trained from scratch, composed of a small number of convolution–ReLU–batch-normalization–pooling blocks, regularization via dropout, and a sigmoid multi-label dense head. These architectures prioritize computational efficiency and serve as both a baseline and a contrast to higher-capacity models.

In parallel, we trained deep networks via transfer learning from *ImageNet*-pretrained backbones. The original classification head was replaced with fully connected layers ending in a K-unit sigmoid output. This scheme was applied to three representative pretrained extractors—MobileNetV2, ResNetV2, and EfficientNetB0—keeping tuning hyperparameters comparable across families.

Aligned with the hypotheses, several experimental factors were activated in a controlled manner. We studied strategies to mitigate class imbalance: per-label weighting of the loss according to prevalence and sampling schemes with moderate positive upweighting, avoiding excessive duplication. In addition, we applied data augmentation policies that preserve radiologic semantics (small rotations, small translations, conservative random crops, and mild brightness/contrast adjustments compatible with radiography), explicitly avoiding left–right flips to preserve laterality. Complementarily, we explored example-mixing techniques that combine pairs of images and targets in a controlled fashion—either by interpolating intensities and labels or by inserting regions from one image into another—to increase the *effective* training diversity.

Training used the *Adam* optimizer, with an initial learning rate of 10^{-3} for from-scratch networks and for the dense head of pretrained models, and 10^{-4} for the fine-tuned feature extractor. Models were trained for up to 50 epochs with per-epoch validation, applying early stopping based on macro AUC-ROC (patience 20) and *ReduceLROnPlateau* (patience 8) to adjust η. This scheme controls overfitting and stabilizes convergence under imbalance.

Typical studies in this area report metrics from a single training run without statistical uncertainty. To adopt a more robust perspective and enable comparative assessment across models, we trained three independent instances per model type and report the mean and standard deviation. These values were used to compare models with Welch's t-test and determine whether differences are significant, as detailed in section 1.

Evaluation followed the protocol established in the Introduction. We report macro-aggregated AUC-ROC and PR-AUC to compare performance across experimental configurations. For comparison with the existing literature, we also report per-label metrics for labels shared across studies. For the model selected as most suitable by discriminative ability, we quantified calibration with the Brier score and illustrated it with reliability diagrams. For operating points, per-label thresholds were set on the validation set to meet high-sensitivity or high-specificity targets, and those thresholds were then applied unchanged to the test set to report the corresponding metrics.

To provide interpretability for the trained algorithms, we generated Grad-CAM maps from the last convolutional layer by weighting activation maps with the class gradients; the result was interpolated and overlaid on the radiograph to produce a heat map, enabling assessment of whether the model focuses on clinically plausible regions for each finding.

The general configuration for the simple from-scratch models used an extractor with layers of 2^{l+3} filters and ReLU activations. After the extractor, *global average pooling* was applied; the dense head consisted of a 256-unit ReLU layer with dropout rate 0.2, followed by a final K-logit layer (one output per label). Training used the weighted binary cross-entropy with logits loss, which operates directly on logits $z_{n,i}$ rather than probabilities; per-label probabilities are obtained via the sigmoid $p_{n,i} = \sigma(z_{n,i}) = \frac{1}{1+e^{-z_{n,i}}}$ for computing the remaining metrics.

The general configuration for the transfer-learning models reported in the next section was as follows. The classification head atop the pretrained extractors consisted of *global average pooling* followed by a large 512-unit ReLU dense layer with dropout 0.2, and a final K-logit layer. In the feature-extractor part, when using pretrained models, the top 90% of layers were unfrozen. The loss was weighted binary cross-entropy with logits.

The numerically stable form of the per-example loss for sample n and label i is

$$\ell_{n,i} \;=\; w_i\Big(\,\max(z_{n,i},0) \;-\; z_{n,i}\,y_{n,i} \;+\; \ln\left(1+e^{-|z_{n,i}|}\right)\Big), \tag{22}$$

and the total loss is averaged over the batch and the K labels,

$$\mathcal{L} \;=\; \frac{1}{N\,K}\sum_{n=1}^{N}\sum_{i=1}^{K}\ell_{n,i}. \tag{23}$$

Per-label weights w_i compensate for imbalance and are set inversely to prevalence π_i (normalized to maintain a comparable scale),

$$w_i \;\propto\; \frac{1}{\pi_i} \qquad \text{with} \qquad \frac{1}{K}\sum_{i=1}^{K} w_i = 1. \tag{24}$$

Data augmentation was performed on the fly via a single pipeline including: random rotation up to $\pm 45°$; height/width zoom in the range $\pm 10\%$; translation in both directions up to 10% of size; random brightness adjustment (maximum deviation 0.4 on rescaled intensity); and addition of uniform noise in $[-0.05, 0.05]$ on preprocessed images with pixel values in $[0, 1]$. These transformations preserve radiologic semantics and aim to increase the effective training diversity.

This configuration follows [29] (except for disabling random horizontal flips in augmentation), which is the strongest published reference in multi-label classification on PadChest and served as the baseline in this work. Nevertheless, it remains subject to change as other configurations with potential improvements are explored.

Two additional loss functions were also considered, both adjusting each example's contribution according to difficulty. In *focal loss*, the cross-entropy is modulated by a factor $(1 - p_t)^\gamma$ that downweights "easy" (already well-classified) examples while maintaining weight on hard ones; an additional factor α_t compensates class imbalance:

$$\ell_{\text{Focal}} = \alpha_t \, (1 - p_t)^\gamma \, (-\log p_t). \tag{25}$$

The class-balanced focal variant replaces α_t with a weight derived from the "effective number" of positive samples for each label, thereby stably increasing the weight of rare classes:

$$\alpha_i^{\text{CB}} = \frac{1 - \beta}{1 - \beta^{n_i}}, \qquad \ell_{\text{CB-Focal}} = \alpha_i^{\text{CB}} \, (1 - p_t)^\gamma \, (-\log p_t). \tag{26}$$

4 Results

This section presents multi-label classification results on PA-projection PadChest radiographs. The radiographic findings selected for the task are the ten most common in the dataset: *Cardiomegaly*, *Aortic elongation*, *Unchanged*, *Scoliosis*, *Costophrenic angle blunting*, *Air trapping*, *Pleural effusion*, *Interstitial pattern*, *Vertebral degenerative changes*, and *Laminar atelectasis*. All metrics reported throughout this section correspond to the test set and are expressed as macro-averaged AUC-ROC and PR-AUC, complemented with per-label results where pertinent.

Comparisons across configurations are organized by model complexity and by experimental factors (e.g., imbalance-handling strategies and data augmentation/mixing), keeping the remaining parameters fixed within each experimental block and consistently using the same evaluation protocol. The 95% confidence intervals are reserved for the final subsection, where we present the model selected as most suitable and complement its evaluation with calibration (Brier score) and illustrative heat maps to provide interpretability.

Lightweight From-Scratch Models vs. Pretrained Models

This comparison includes shallow networks trained from scratch and higher-capacity pretrained architectures. We denote by CNN-X a lightweight, simple model built from scratch and initialized with random weights, with X convolutional blocks (each block: convolution–activation–batch normalization–*pooling*); pretrained transfer-learning models are referred to by their standard names from *ImageNet*.

Table 1: Performance of lightweight CNNs trained from scratch versus pretrained architectures across model families. Metrics: macro-averaged AUC-ROC and PR-AUC. "Parameters (M)" indicates the total number of model parameters (in millions).

Model	Regime	Parameters (M)	AUC-ROC	PR-AUC
CNN-3	From scratch	0.04	0.600 ± 0.030	0.130 ± 0.020
CNN-4	From scratch	0.13	0.640 ± 0.030	0.150 ± 0.020
CNN-5	From scratch	0.46	0.670 ± 0.030	0.160 ± 0.018
CNN-6	From scratch	1.71	0.670 ± 0.030	0.160 ± 0.018
MobileNetV2	Fine-tuning (TL)	3.40	0.734 ± 0.009	0.182 ± 0.010
ResNet50V2	Fine-tuning (TL)	25.60	0.741 ± 0.007	0.187 ± 0.008
EfficientNetB0	Fine-tuning (TL)	5.30	0.739 ± 0.006	0.185 ± 0.008

As summarized in Table 1, aggregate performance increases with depth for lightweight from-scratch CNNs, but a clear plateau appears between CNN-5 and CNN-6. In contrast, all pretrained architectures outperform the lightweight from-scratch ones. Among the pretrained models, MobileNetV2 is slightly below the other two; ResNet50V2 and EfficientNetB0 show no significant difference (AUC-ROC: $|t| = 0.38 < t^* = 2.79$, $p = 0.72$; PR-AUC: $|t| = 0.31 < t^* = 2.78$, $p = 0.76$).

EfficientNetB0 Fine-Tuning: Unfrozen Layers

From this point on we fix EfficientNetB0 as the feature extractor (given its lower computational cost compared with ResNet50V2) and compare different percentages of unfrozen layers during fine-tuning. The percentage indicates the top fraction of the backbone made trainable (the head is always trained); the rest of the protocol remains unchanged.

Table 2: EfficientNetB0 performance at different percentages of unfrozen layers. Metrics: macro-averaged AUC-ROC and PR-AUC. "Parameters (M)" shows the number of *trainable* parameters (in millions).

Unfrozen fraction	Regime	Parameters (M)	AUC-ROC	PR-AUC
0 %	Head only	0.40	0.610 ± 0.012	0.110 ± 0.010
25 %	Partial fine-tuning	1.50	0.665 ± 0.011	0.140 ± 0.011
50 %	Partial fine-tuning	2.90	0.705 ± 0.010	0.165 ± 0.010
75 %	Partial fine-tuning	4.20	0.730 ± 0.008	0.178 ± 0.009
90 %	Extensive fine-tuning	5.00	0.739 ± 0.006	0.185 ± 0.008
100 %	Full fine-tuning	5.30	0.736 ± 0.007	0.183 ± 0.009

Overall, as shown in Table 2, performance increases markedly with the fraction of trainable layers up to ∼90 %; fully unfreezing does not yield further gains and shows a slight, non-significant drop (AUC-ROC: $|t| = 0.56 < t^* = 2.80$, $p = 0.60$; PR-AUC: $|t| = 0.29 < t^* = 2.79$, $p = 0.78$). Henceforth, we proceed with the 90 % model as it offers the best performance at a reasonable cost.

Weight Initialization: Pretraining vs. Random

With EfficientNetB0 fixed as the backbone, we compare the effect of weight initialization. We contrast fine-tuning from *ImageNet*-pretrained weights (with 90 % of the extractor unfrozen) against random initialization, in which case all layers are trained.

Table 3: EfficientNetB0 performance with pretrained vs. random initialization. Metrics: macro-averaged AUC-ROC and PR-AUC. "Parameters (M)" indicates the total number of trainable parameters (in millions).

Configuration	Regime	Parameters (M)	AUC-ROC	PR-AUC
Pretrained	Fine-tuning	5.00	0.739 ± 0.006	0.185 ± 0.008
Random	Full training	5.30	0.701 ± 0.010	0.146 ± 0.010

As shown in Table 3, performance with pretrained initialization is significantly superior to that with random weights (AUC-ROC: $|t| = 5.64 > t^* = 3.07$, $p = 0.006$; PR-AUC: $|t| = 5.27 > t^* = 2.80$, $p = 0.006$).

Classifier Structure (Dense Head)

We compare different architectures for the classification head. The notation "$u_1/\delta_1 \to u_2/\delta_2$" denotes two sequential dense layers with u_1 and u_2 units and dropout rates δ_1 and δ_2, respectively; when a single pair appears, it indicates a single dense layer.

Table 4: EfficientNetB0 performance across classification-head structures. Metrics: macro-averaged AUC-ROC and PR-AUC. Notation: "u/δ" denotes units/dropout per layer.

Structure (units / dropout)	AUC-ROC	PR-AUC
512 / 0.20	0.739 ± 0.006	0.185 ± 0.008
512 / 0.40 $\to$ 256 / 0.30	0.736 ± 0.007	0.181 ± 0.009
256 / 0.30 $\to$ 128 / 0.30	0.753 ± 0.006	0.204 ± 0.008
128 / 0.20 $\to$ 64 / 0.20	0.734 ± 0.010	0.176 ± 0.011

As shown in Table 4, the 256/0.30 $\to$ 128/0.30 configuration attains the best averages. The difference versus 512/0.20 is statistically significant for both AUC-ROC ($|t| = 2.86 > t^* = 2.78$, $p = 0.046$) and PR-AUC ($|t| = 2.91 > t^* = 2.78$, $p = 0.042$). Hereafter, we retain the 256/0.30 $\to$ 128/0.30 head as the classifier.

With the network configuration finalized (base architecture, percentage of trainable layers, weight initialization, and head structure), we now focus on the training setup to address class imbalance. Specifically, we compare different loss functions, sampling strategies, and data augmentation techniques.

Loss Functions

With the architecture fixed (EfficientNetB0 pretrained on ImageNet with 90 % of layers unfrozen and a 256/0.30 $\to$ 128/0.30 head), we compare three loss functions designed to mitigate label imbalance: weighted binary cross-entropy with logits (weighted BCE with logits), focal loss, and class-balanced focal loss.

Table 5: Performance across loss functions. Metrics: macro-averaged AUC-ROC and PR-AUC.

Loss function	AUC-ROC	PR-AUC
Weighted BCE with logits	0.753 ± 0.006	0.204 ± 0.008
Focal (α=0.20, γ=2.5)	0.734 ± 0.010	0.180 ± 0.011
Class-balanced focal	0.738 ± 0.006	0.185 ± 0.007

As shown in Table 5, training with weighted BCE with logits attains the highest averages and is superior to the next best (class-balanced focal): AUC-ROC $|t| = 3.06 > t^* = 2.78$, $p = 0.038$; PR-AUC $|t| = 3.10 > t^* = 2.80$, $p = 0.037$.

Minority-Class Oversampling

We next evaluate the effect of applying moderate positive oversampling versus not applying it, keeping the previous configuration fixed.

Table 6: Performance with and without moderate oversampling of minority classes. Metrics: macro-averaged AUC-ROC and PR-AUC.

Configuration	AUC-ROC	PR-AUC
Without oversampling	0.753 ± 0.006	0.204 ± 0.008
With moderate oversampling	0.767 ± 0.006	0.223 ± 0.008

As shown in Table 6, moderate oversampling yields higher averages. The improvement is marginally significant relative to not oversampling (AUC-ROC: $|t| = 2.86 > t^* = 2.78$, $p = 0.046$; PR-AUC: $|t| = 2.91 > t^* = 2.78$, $p = 0.042$). Hence, we adopt it as the default.

With the above decisions, we consider the following configuration as our intermediate model: EfficientNetB0 pretrained on *ImageNet*, with 90% of the backbone unfrozen, a 256 / 0.30 $\rightarrow$ 128 / 0.30 dense head, weighted BCE-with-logits loss, and moderate positive oversampling. We use this model as an internal reference; the per-label comparison against the baseline of Liz et al. [29] is presented later in Table 9.

Data Augmentation Techniques

In this subsection we analyze the effect of different augmentation strategies. The basic scheme includes the augmentation described in the Materials and Methods section; on top of it we incorporate example-mixing techniques: interpolation of image–label pairs (*MixUp* [30]) and insertion of a patch from one image into another with a composite label (*CutMix* [31]). Both processes occur at random, and the example labels are updated in proportion to the mix to reflect the new image content.

Table 7: Performance under different data augmentation schemes. Metrics: macro-averaged AUC-ROC and PR-AUC.

Augmentation	AUC-ROC	PR-AUC
Basic	0.767 ± 0.006	0.223 ± 0.008
Basic + *MixUp* (α=0.40)	0.764 ± 0.009	0.221 ± 0.010
Basic + *CutMix*	0.780 ± 0.005	0.241 ± 0.007

As shown in Table 7, *Basic + CutMix* achieves the highest averages and is significantly superior to the *Basic* scheme (AUC-ROC: $|t| = 2.88 > t^* = 2.81$, $p = 0.045$; PR-AUC: $|t| = 2.93 > t^* = 2.80$, $p = 0.040$). By contrast, *MixUp* does not yield improvements and slightly worsens *Basic* on both metrics.

Final Comparison Across Architectures

We apply the final training setup to the three architectures considered (MobileNetV2, ResNet50V2, and EfficientNetB0) to assess performance under the new measures: 256/0.30 → 128/0.30 head, moderate oversampling, and the *Basic + CutMix* augmentation scheme.

Table 8: Compared performance of architectures under the final configuration. Metrics: macro-averaged AUC-ROC and PR-AUC. "Parameters (M)" indicates the total number of model parameters (in millions).

Model	Parameters (M)	AUC-ROC	PR-AUC
MobileNetV2	3.40	0.778 ± 0.006	0.238 ± 0.008
ResNet50V2	25.60	0.781 ± 0.005	0.242 ± 0.007
EfficientNetB0	5.30	0.780 ± 0.005	0.241 ± 0.007

As shown in Table 8, the three architectures perform very similarly. Taking MobileNetV2 as the reference, differences versus ResNet50V2 and EfficientNetB0 are not statistically significant for either AUC-ROC or PR-AUC. For AUC-ROC, versus ResNet50V2 we obtain $|t| = 0.67$, $t^* = 2.81$, $p = 0.54$; versus EfficientNetB0, $|t| = 0.44$, $t^* = 2.79$, $p = 0.68$. For PR-AUC, versus ResNet50V2 we obtain $|t| = 0.65$, $t^* = 2.80$, $p = 0.54$; versus EfficientNetB0, $|t| = 0.49$, $t^* = 2.78$, $p = 0.64$. In all cases $|t| < t^*$ and $p > 0.05$. Given its lower computational cost, we retain MobileNetV2 as the preferred option without loss of performance.

Per-Label Comparison with Liz et al. [29]

To position our models against prior work on the same dataset, we compare with Liz et al. [29]. That study trains five backbone architectures (DenseNet-201, EfficientNet-B0, Inception, Inception-ResNet, and Xception) and also reports several ensembles. Here we restrict ourselves to their *EfficientNet* column trained with segmentation-based cropping and data augmentation (Table 8 in their paper), as it is most comparable to our pipeline and avoids mixing ensemble results with single-network results.

Because Liz et al. [29] evaluates a broader label set (including differential diagnoses), we restrict the comparison to the intersection with our ten trained labels. Their work reports only the threshold-free AUC-ROC and an F1 score at an undocumented operating point. Therefore—consistent with our aim of giving clinical meaning to model decisions—we compare only AUC-ROC here and later report threshold-dependent metrics with clinical rationale.

Table 9 lists, for each label, the AUC-ROC reported by *Liz et al.* and the difference Δ with two of our systems: (i) an intermediate model: EfficientNet-B0 with the configuration described up to the minority-class oversampling subsection; and (ii) the selected final model: MobileNetV2 with *CutMix*. We define $\Delta = \mathrm{AUC}_{\mathrm{ours}} - \mathrm{AUC}_{\mathrm{Liz}}$, so $\Delta > 0$ indicates

improvement over their result.

Table 9: Per-label performance on PadChest compared with the reference Liz et al. [29]. The "Liz, EffNet" column comes from Table 8 of Liz et al. [29]. "Δ" is defined as $\Delta = \mathrm{AUC}_{\mathrm{ours}} - \mathrm{AUC}_{\mathrm{Liz}}$ (positive indicates improvement).

Label	**Liz, EffNet**	**Δ (intermediate)**	**Δ (final)**
Cardiomegaly	0.907	-0.006 ± 0.011	$+0.002 \pm 0.011$
Aortic elongation	0.846	-0.013 ± 0.011	$+0.024 \pm 0.012$
Pleural effusion	0.951	-0.016 ± 0.012	-0.055 ± 0.012
Scoliosis	0.712	$+0.007 \pm 0.012$	$+0.084 \pm 0.013$
Interstitial pattern	0.795	-0.004 ± 0.011	-0.003 ± 0.012
Costophrenic angle b.	0.845	-0.022 ± 0.012	-0.121 ± 0.014
Air trapping	0.687	$+0.001 \pm 0.012$	$+0.037 \pm 0.013$
Unchanged	0.614	-0.018 ± 0.013	-0.039 ± 0.013
Vertebral degenerative c.	0.721	-0.002 ± 0.011	$+0.053 \pm 0.012$
Laminar atelectasis	0.806	-0.014 ± 0.013	-0.085 ± 0.014

As shown in Table 9, the intermediate model (EfficientNetB0) performs slightly below Liz et al. [29] for most labels such as *Aortic elongation* and *Pleural effusion*, and is virtually tied on a few labels such as *Scoliosis* and *Air trapping*, considering variability across seeds.

The final model (MobileNetV2+CutMix) exhibits greater per-label variability—with more pronounced ups and downs—but overall surpasses both Liz et al. [29] and the intermediate model. Relative to Liz, it stands out on structural findings such as *Scoliosis* and *Vertebral degenerative changes*, with gains also in *Air trapping*, while it drops on *Pleural effusion*, *Costophrenic angle blunting*, *Laminar atelectasis*, and *Unchanged*. Compared with the intermediate model, its largest gains are concentrated precisely on the aforementioned structural labels, tipping the overall balance in favor of the final model.

Calibration and Operating Points

Table 10 reports calibration (Brier) and two per-label operating points: (i) specificity at fixed sensitivity of 95%, and (ii) sensitivity at fixed specificity of 95%. Thresholds were set on the validation set and applied unchanged to the test set.

Table 10: Calibration and operating points of the final model, per label; Brier and two operating points are reported. Metrics: Brier (lower is better), "Spec@Sens=95%" (specificity at fixed sensitivity of 95%), and "Sens@Spec=95%" (sensitivity at fixed specificity of 95%).

Label	**Brier**	**Spec@Sens=95%**	**Sens@Spec=95%**
Cardiomegaly	0.181 ± 0.012	0.610 ± 0.040	0.500 ± 0.040
Aortic elongation	0.149 ± 0.013	0.530 ± 0.030	0.350 ± 0.040
Unchanged	0.380 ± 0.020	0.060 ± 0.020	0.090 ± 0.020
Scoliosis	0.210 ± 0.014	0.180 ± 0.030	0.330 ± 0.040
Costophrenic angle b.	0.168 ± 0.013	0.150 ± 0.030	0.240 ± 0.030
Air trapping	0.210 ± 0.014	0.140 ± 0.030	0.180 ± 0.030
Pleural effusion	0.034 ± 0.006	0.480 ± 0.030	0.730 ± 0.030
Interstitial pattern	0.073 ± 0.010	0.220 ± 0.030	0.330 ± 0.030
Vertebral degenerative c.	0.070 ± 0.010	0.170 ± 0.030	0.240 ± 0.030
Laminar atelectasis	0.065 ± 0.010	0.160 ± 0.030	0.260 ± 0.030
Macro	0.154 ± 0.015	0.270 ± 0.030	0.320 ± 0.040

To visually illustrate the operating points, Figure 4 shows, for four representative findings, the ROC curves with both thresholds marked: the point at Sens=95% (which yields the "Spec@Sens=95%" column of Table 10) and the point at Spec=95% (which yields the "Sens@Spec=95%" column).

Analogously, Figure 5 depicts, for the same labels, the test-set probability histograms (score distributions for positives and negatives) that help assess model calibration quality.

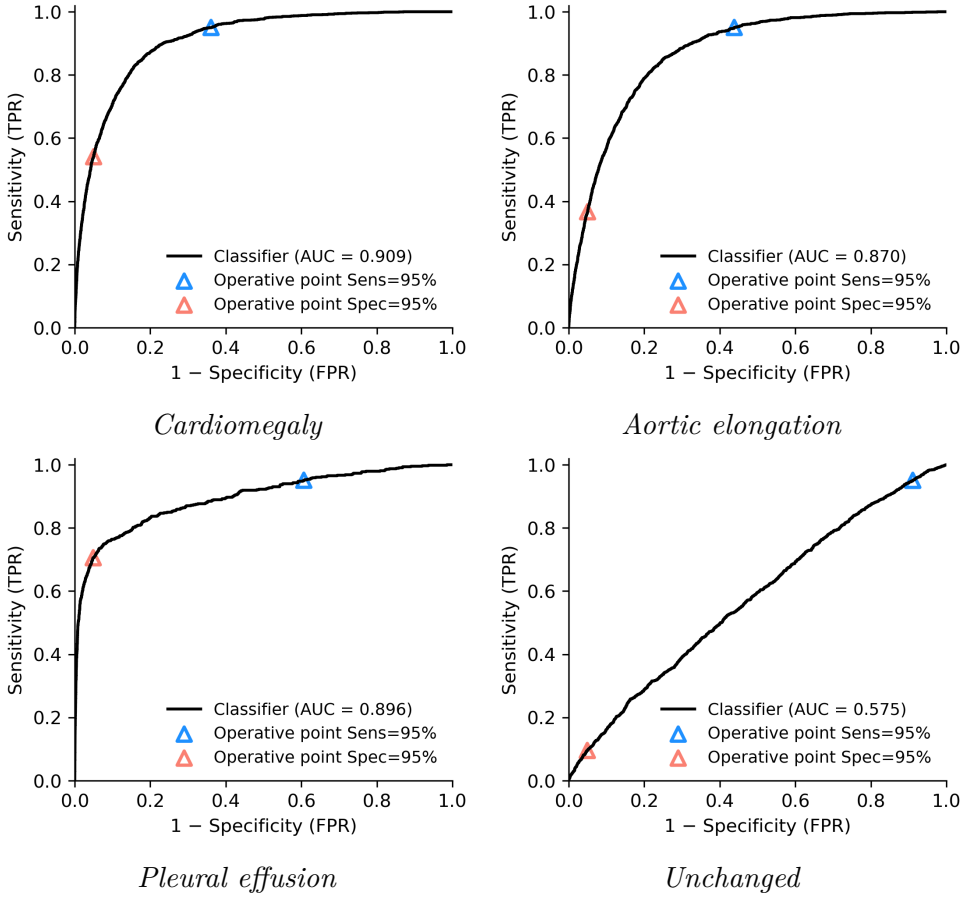


Figure 4: ROC curves with operating points (Sens=95 % and Spec=95 %) for four representative findings.

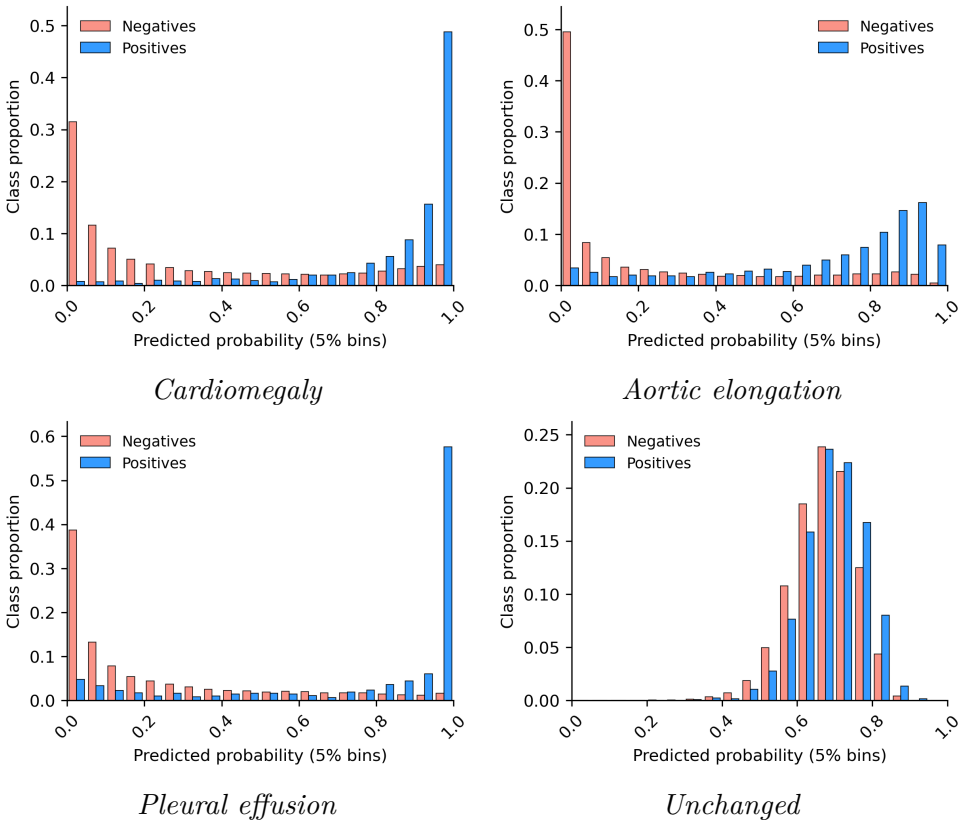
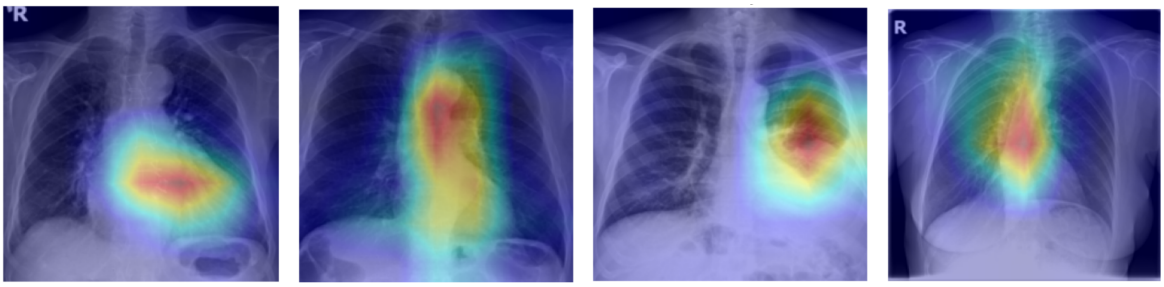


Figure 5: Per-label probability histograms for four representative findings. Positives (blue) vs. negatives (orange). Relative frequency of the predicted probability by class.



Cardiomegaly *Aortic elongation* *Pleural effusion* *Scoliosis*

Figure 6: Grad-CAM heat maps overlaid on the radiograph for four examples with a single positive label and model probability > 0.9. Reddish tones indicate higher contribution to the prediction, bluish tones lower contribution.

Interpretability: Heat Maps

To visualize which regions support the predictions, Figure 6 shows Grad-CAM maps overlaid on the radiograph for four representative findings. Each example depicts an image with a single positive label and model probability above 0.9 for that label; areas contributing most to the decision are highlighted, facilitating assessment of the anatomical plausibility of the prediction.

5 Discussion

We now discuss the results presented in the previous section. We analyze performance differences across models and compare them with the literature, both in general computer vision and in chest radiography in particular.

Model Architecture

The results in Table 1 show that lightweight from-scratch CNNs surpass the random-classifier threshold (AUC-ROC > 0.5) but remain far from acceptable performance on this multi-label problem. Although performance improves with depth, the gains plateau beyond five blocks, indicating insufficient capacity to capture the radiographic variability of the dataset. In contrast, pretrained reference architectures (MobileNetV2, ResNet50V2, EfficientNetB0) achieve substantial improvements even under a suboptimal configuration. This behavior aligns with the literature: complex visual tasks benefit from architectures with specific mechanisms—residual blocks, depthwise separable convolutions, and squeeze-and-excitation attention—that increase expressivity without exploding computational cost [7, 24, 26, 32]. Hence, they are chosen as the basis for the remainder of the experiments.

When analyzing the fraction of unfrozen layers during fine-tuning (Table 2), a clear pattern emerges: allowing a larger portion of the backbone to be retrained increases performance up to roughly 90%, while fully unfreezing yields no further gains and can introduce slight instability. This is expected because early layers typically learn generic features (edges, textures) useful across domains, whereas later layers capture target-specific characteristics; consequently, prioritizing updates to upper layers is often more effective than modifying the entire backbone [33]. In medical imaging—where the domain differs substantially from *ImageNet*—adjusting a large fraction of the extractor is especially relevant to narrow the domain gap [34].

The comparison between pretrained and random initialization (Table 3) confirms that *ImageNet* pretraining offers a consistent advantage under our experimental regimen. This accords with evidence that (i) representations learned on *ImageNet* often transfer with measurable benefits, and (ii) better source performance tends to translate into better transfer, particularly when target data are limited [34]. While, with sufficient epochs and resources, training from scratch may approach pretrained performance on certain tasks, enforcing a homogeneous epoch budget across configurations more fairly exposes the efficiency of transfer in this context.

Finally, regarding the model's architectural configuration, the study of the classification head (Table 4) shows that a two-layer, moderately wide head ($256/0.30 \rightarrow 128/0.30$) outperforms both wider and narrower alternatives. This suggests a favorable balance between representational capacity and regularization (via dropout) after global average pooling. It also constitutes a key difference from Liz et al. [29], where a single 512-unit

layer with 0.20 dropout is used. Because the head operates on enriched representations, small changes in width and depth can affect the smoothness of the decision boundary and the resulting calibration. The observed superiority should thus be interpreted as specific to the present pipeline and label set; therefore, these head-design results are not directly comparable with prior studies.

Strategies to Counter Class Imbalance

The experiments in this subsection were designed to mitigate the strong imbalance between positive and negative cases characteristic of PadChest. We evaluate three complementary levers: (i) loss functions robust to imbalance; (ii) moderate oversampling of minority positives combined with on-the-fly augmentation; and (iii) example-mixing techniques. The goal is to determine which combination consistently improves discrimination (AUC-ROC, PR-AUC).

In Table 5, weighted binary cross-entropy with logits (BCE with per-label prevalence weights) attains the best averages and significantly outperforms focal variants. This aligns with the widespread use of BCE in multi-label classification—its logits form is numerically stable and, when per-label weighting is applied, directly compensates frequency imbalances—thus maximizing likelihood and inducing good ranking for threshold-free metrics [35]. By contrast, *focal loss* downweights "easy" examples to focus on hard ones [36], and the class-balanced variant reweights via the "effective number" of samples to assist rare classes [37]. However, both introduce sensitivity to hyperparameters (γ, α, β) that require careful tuning—particularly challenging in a multi-label setting with large prevalence disparities. Consequently, they did not yield a net advantage over a well-calibrated weighted BCE. In short, adding per-label weights to BCE proved a simpler and more effective solution for this dataset.

Table 6 shows that moderately re-exposing minority positives—always with on-the-fly augmentation—significantly improves both AUC-ROC and PR-AUC relative to not oversampling. This behavior is consistent with prior results: when classes are rare, increasing the sampling probability of positives can reduce gradient bias toward the majority class, especially when combined with regularization and augmentations that prevent memorization [38]. In our case, weighted BCE handled varying class prevalences well, and oversampling acted as a complementary mechanism that increased coverage without inflating false positives. This departs from Liz et al. [29], which does not use explicit oversampling; here we observe that a moderate, controlled boost yields a reproducible benefit without evident artifacts.

Up to this point, we have explored the typical configuration described in reference works on computer vision for chest radiography. The model selected as best—EfficientNetB0 pretrained on *ImageNet*, fine-tuned with 90 % of the backbone unfrozen, a 256/0.30 $\rightarrow$ 128/0.30 dense head, and weighted BCE with logits, plus moderate positive oversampling—differs slightly from Liz et al. [29]. The most relevant differences are: (i) a two-

layer, moderately wide head instead of a single wider layer; and (ii) explicit moderate minority oversampling. The direct comparison with Liz et al. [29] is given in Table 9 and discussed later.

On top of the basic set of radiography-compatible geometric and photometric augmentations, we compared *MixUp* and *CutMix*. As summarized in Table 7, *MixUp* did not help: linear intensity interpolation tends to blur low-contrast, high–spatial-frequency radiographic structures (pleural edges, cardiac contours), diluting key local cues. In contrast, *CutMix* replaces localized regions between images while preserving background texture and anatomy, which encourages attention to discriminative zones and acts as an effective spatial regularizer, yielding significant gains on both metrics. In domains where anatomical localization matters, preserving local sharpness and structural coherence appears more beneficial than global intensity interpolation.

The final cross-architecture comparison (Table 8) shows that the training-regime improvements—256 / 0.30 → 128 / 0.30 head, moderate oversampling, and *Basic + CutMix* augmentation—benefit MobileNetV2, ResNet50V2, and EfficientNetB0 almost uniformly, to the point that differences in AUC-ROC and PR-AUC are not statistically significant. This pattern suggests that most of the progress stems from the training procedure rather than peculiarities of any particular family. Consequently, and given its lower computational cost (3.4 M parameters vs. 25.6 M for ResNet50V2, with indistinguishable performance), we retain MobileNetV2 as the preferred option for deployment due to its better precision–efficiency trade-off without sacrificing discriminative capacity.

Comparison with Liz et al. [29]

In the direct per-label comparison with Liz et al. [29] (Table 9), our final model (MobileNetV2+*CutMix*) exhibits a differential pattern consistent with the design choices: it clearly improves on structurally or morphologically driven findings (*Scoliosis*, *Vertebral degenerative changes*), whereas it underperforms on more "diffuse" or smooth-gradient findings (*Pleural effusion*, *Costophrenic angle blunting*, *Laminar atelectasis*, *Unchanged*). We interpret these divergences as the combined effect of (i) a two-layer, moderately wide dense head that smooths the decision boundary and favors stable anatomical patterns over subtle textures; (ii) moderate minority oversampling, which increases coverage of rare positives without aggressive duplication; and (iii) *CutMix*, which preserves local sharpness and encourages attention to discriminative regions—particularly advantageous for structural labels [31].

By contrast, the intermediate model (EfficientNetB0) does not significantly outperform the model reported by Liz et al. [29] on any label, although it delivers very similar performance on most labels—an expected outcome given the similarity between training setups. It is also worth noting that Liz et al. [29] trains with substantially higher epoch limits than those used here, which may yield slight performance gains at the cost of a considerable increase in training time.

Calibration, Operating Points, and Heat Maps

The Brier scores in Table 10 indicate overall adequate calibration (macro 0.154 ± 0.015), with values below 0.20 for most labels, suggesting that predicted probabilities reasonably reflect observed frequencies [39]. The exception is *Unchanged* (Brier 0.380 ± 0.020), consistent with its low AUC-ROC and with the nature of the label: it does not describe a radiographic finding per se but an assessment (change vs. no change) that cannot be recovered from a single PA image. At the opposite end, *Pleural effusion* shows outstanding calibration (Brier 0.034 ± 0.006), in line with its best discrimination.

Operating points fixed at 95 % reveal the trade-off between false positives and false negatives: enforcing Sens=95 % prioritizes avoiding false negatives at the cost of reduced specificity (e.g., *Cardiomegaly*: Spec 0.61 ± 0.04; *Pleural effusion*: 0.48 ± 0.03), whereas fixing Spec=95 % contains the false-positive rate while accepting sensitivity losses (e.g., *Cardiomegaly*: Sens 0.50 ± 0.04; *Aortic elongation*: 0.35 ± 0.04). In a clinical pipeline, the former is more useful for early screening (rule-out: high sensitivity, allowing subsequent verification), and the latter for late confirmation (high specificity to minimize unwarranted alarms). In light of these results, the model would function better as late confirmation for *Pleural effusion* (high Sens at Spec=95%) and as early screening for *Cardiomegaly* and *Aortic elongation* (Sens=95% with moderate specificities that help prioritize studies); by contrast, for low-contrast or weak-semantics labels such as *Unchanged*, operational use with fixed thresholds is not recommended.

Taken together, the pairs obtained at the 95 % settings (macro: Spec@Sens=95 % ≈ 0.27, Sens@Spec=95 % ≈ 0.32) are modest and consistent with the reduced macro PR-AUC (≤ 0.24). Under severe imbalance, the ROC curve can be relatively optimistic because FPR is damped by the abundance of negatives, whereas the Precision–Recall curve directly penalizes false predictions.

The ROC curves in Figure 4 reflect these patterns: points near the upper-left corner—ideal of high sensitivity and low FPR—are seen "reasonably" for *Cardiomegaly*, *Aortic elongation*, and *Pleural effusion*, in line with their strong discrimination and calibration; by contrast, *Unchanged* lies far from that region, confirming its poor separability. The histograms in Figure 5 tell the same story from another angle: clear separation of score distributions for the first three labels versus an almost total overlap for *Unchanged*, which prevents setting thresholds with a good balance. This contrast reinforces the idea that *Unchanged* should be approached with longitudinal methods (image pairs or temporal information) rather than as a finding detectable in a single radiograph.

The maps in Figure 6 show anatomically plausible activation foci consistent with the positive label in each example: for *Cardiomegaly*, the heat concentrates on the cardiac silhouette; for *Aortic elongation*, the mediastinal/aortic contour is highlighted; for *Pleural effusion*, the affected lung base predominates in the corresponding case; and for *Scoliosis*, the vertebral axis is emphasized. This correspondence supports that the model leverages

relevant regions for its decisions rather than spurious signals, in line with Grad-CAM's aim of weighting activations by gradients in the last convolutional layer [14].

6 Conclusions

This work addressed multi-label classification of findings in chest radiographs (PadChest, PA projection) using convolutional networks and a comparative experimental protocol. We analyzed both lightweight architectures trained from scratch and pretrained models via transfer learning, contrasting architectural configurations and, systematically, training factors relevant for imbalanced data.

Methodologically, we followed a controlled comparison strategy across configurations: choice of architectures (MobileNetV2, ResNet50V2, EfficientNetB0), fraction of unfrozen layers during fine-tuning, classifier-head structure, loss function, moderate oversampling, and data augmentation/mixing schemes. Each decision was evaluated with threshold-free metrics and repeated across multiple seeds to estimate training variability.

The results show that performance depends more on the training procedure than on the specific backbone family. In particular: (i) fine-tuning with ImageNet pretraining and 90% of the backbone unfrozen clearly outperforms training from scratch; (ii) a moderately sized two-layer head (256/0.30 → 128/0.30) improves over both wider and narrower alternatives; (iii) weighted BCE with logits is more effective and stable than focal variants in this setting; (iv) moderate positive oversampling provides an additional gain; and (v) CutMix, on top of a basic augmentation scheme, significantly increases AUC-ROC and PR-AUC. With the final configuration, the three architectures achieve very similar performance (Table 8); therefore, MobileNetV2 is selected as the preferred model for its better precision–efficiency trade-off.

In comparison with prior work on PadChest, we provide a direct reference against Liz et al. [29]. Our intermediate model (EfficientNetB0) is overall close but somewhat below, whereas the final model (MobileNetV2 + CutMix) delivers notable improvements on structurally driven labels (e.g., *Scoliosis*, *Vertebral degenerative changes*) and declines on more diffuse findings (e.g., *Pleural effusion*, *Laminar atelectasis*), yielding an overall favorable balance (Table 9).

To the best of our knowledge, this is the first study in this setting (multi-label chest radiograph classification) to jointly report PR-AUC, an explicit calibration measure (Brier score), and the standard deviation due to training variability. We also present clinically useful operating points—Spec@Sens=95% and Sens@Spec=95%—together with Grad-CAM maps that provide anatomical interpretability. These elements help estimate system behavior in high-sensitivity screening or high-specificity confirmation scenarios and enhance model transparency.

Limitations and future work: (1) Scope expansion: the study is restricted to 10 labels and the PA projection; natural extensions include broadening the set of findings and incorporating available PadChest metadata (e.g., age, sex, and other patient variables) as covariates or via multimodal heads, as well as exploring pretraining on domain-specific

chest X-ray datasets. (2) External validation: evaluate the model on an independent, comparable dataset (another analogous public collection) and, if needed, consider recalibration for the new domain. (3) Ensembling and uncertainty estimation: investigate lightweight ensembles (averaging multiple architectures) and methods to quantify uncertainty to capture variability due to both the training procedure and the finite size of the evaluation set, reporting per-label confidence intervals and curves. (4) New architectures: beyond CNNs, newer vision backbones such as *vision transformers* are rapidly advancing and merit exploration.

Overall, the work establishes a solid, reproducible, and clinically oriented baseline for multi-label classification on PadChest chest radiographs, identifying training decisions that generalize across architectures and providing metrics and tools (PR-AUC, calibration, variability, operating points, and interpretability) that facilitate evaluation and potential integration into clinical practice.

References

[1] T. M. Mitchell. *Machine Learning*. New York: McGraw-Hill, 1997.

[2] Y. Bengio et al. "Representation learning: A review and new perspectives". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (2013), pp. 1798–1828. DOI: 10.1109/TPAMI.2013.50.

[3] Y. LeCun et al. "Deep learning". In: *Nature* 521.7553 (2015), pp. 436–444. DOI: 10.1038/nature14539.

[4] G. Litjens et al. "A survey on deep learning in medical image analysis". In: *Medical Image Analysis* 42 (2017), pp. 60–88. DOI: 10.1016/j.media.2017.07.005.

[5] Y. LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. DOI: 10.1109/5.726791.

[6] A. Krizhevsky et al. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2012.

[7] K. He et al. "Deep Residual Learning for Image Recognition". In: *CVPR*. 2016. DOI: 10.1109/CVPR.2016.90.

[8] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[9] S. Ioffe and C. Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*. Vol. 37. JMLR Workshop and Conference Proceedings. 2015, pp. 448–456.

[10] A. Zafar et al. "A Comparison of Pooling Methods for Convolutional Neural Networks". In: *Applied Sciences* 12.17 (2022), p. 8643. DOI: 10.3390/app12178643.

[11] M. Lin et al. "Network In Network". In: *International Conference on Learning Representations (ICLR)*. 2014. arXiv: 1312.4400 [cs.NE].

[12] S. Skansi. *Introduction to Deep Learning: From Logical Calculus to Artificial Intelligence*. Undergraduate Topics in Computer Science. Cham: Springer, 2018. DOI: 10.1007/978-3-319-73004-2. URL: https://link.springer.com/book/10.1007/978-3-319-73004-2.

[13] D. M. W. Powers. "Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness & Correlation". In: *Journal of Machine Learning Technologies* 2.1 (2011), pp. 37–63.

[14] R. R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *International Journal of Computer Vision* 128.2 (2019), pp. 336–359. DOI: 10.1007/s11263-019-01228-7. URL: http://dx.doi.org/10.1007/s11263-019-01228-7.

[15] G. W. Brier. "Verification of Forecasts Expressed in Terms of Probability". In: *Monthly Weather Review* 78.1 (1950), pp. 1–3. DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[16] T. Fawcett. “An introduction to ROC analysis”. In: *Pattern Recognition Letters* 27.8 (2006), pp. 861–874. DOI: 10.1016/j.patrec.2005.10.010.

[17] J. Davis and M. Goadrich. “The relationship between Precision-Recall and ROC curves”. In: *Proceedings of the 23rd International Conference on Machine Learning*. ACM, 2006, pp. 233–240. DOI: 10.1145/1143844.1143874.

[18] B. L. Welch. “The Generalization of Student’s Problem when Several Different Population Variances are Involved”. In: *Biometrika* 34.1-2 (1947), pp. 28–35. DOI: 10.1093/biomet/34.1-2.28.

[19] G. Casella and R. L. Berger. *Statistical Inference*. 2nd ed. Pacific Grove, CA: Duxbury / Thomson Learning, 2002.

[20] A. Bustos et al. “PadChest: A large chest x-ray image dataset with multi-label annotated reports”. In: *Medical Image Analysis* 66 (2020), p. 101797. DOI: 10.1016/j.media.2020.101797. URL: https://doi.org/10.1016/j.media.2020.101797.

[21] M. Abadi et al. “TensorFlow: A System for Large-Scale Machine Learning”. In: *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*. 2016, pp. 265–283.

[22] F. Chollet et al. *Keras*. https://keras.io/. 2015.

[23] O. Russakovsky et al. “ImageNet Large Scale Visual Recognition Challenge”. In: *International Journal of Computer Vision* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[24] M. Sandler et al. “MobileNetV2: Inverted Residuals and Linear Bottlenecks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 4510–4520. DOI: 10.1109/CVPR.2018.00474.

[25] K. He et al. “Identity Mappings in Deep Residual Networks”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2016, pp. 630–645. DOI: 10.1007/978-3-319-46493-0_38.

[26] M. Tan and Q. V. Le. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”. In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*. 2019, pp. 6105–6114.

[27] J. P. Cohen, J. D. Viviano, et al. “TorchXRayVision: A library of chest X-ray datasets and models”. In: *Proceedings of the Machine Learning Research*. Vol. 172. 2022. URL: https://mlmed.org/torchxrayvision/.

[28] H. Zhao et al. “Pyramid Scene Parsing Network”. In: *CVPR*. 2017.

[29] H. Liz et al. “Deep learning for understanding multilabel imbalanced Chest X-ray datasets”. In: *Future Generation Computer Systems* 144 (2023), pp. 291–306. DOI: 10.1016/j.future.2023.03.005. URL: https://doi.org/10.1016/j.future.2023.03.005.

[30] H. Zhang et al. “mixup: Beyond Empirical Risk Minimization”. In: *International Conference on Learning Representations (ICLR)*. 2018. URL: https://openreview.net/forum?id=r1Ddp1-Rb.

[31] S. Yun et al. “CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 6023–6032. DOI: 10.1109/ICCV.2019.00612.

[32] J. Hu et al. “Squeeze-and-Excitation Networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 7132–7141. DOI: 10.1109/CVPR.2018.00745.

[33] J. Yosinski et al. “How Transferable are Features in Deep Neural Networks?” In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2014, pp. 3320–3328.

[34] M. Raghu et al. “Transfusion: Understanding Transfer Learning for Medical Imaging”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019, pp. 3342–3352.

[35] Z. Zhang et al. “Classification with Deep Neural Networks and Logistic Loss”. In: *Journal of Machine Learning Research* 25.125 (2024), pp. 1–117. arXiv: 2307.16792. URL: https://www.jmlr.org/papers/v25/22-0049.html.

[36] T.-Y. Lin et al. “Focal Loss for Dense Object Detection”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2980–2988. DOI: 10.1109/ICCV.2017.324.

[37] Y. Cui et al. “Class-Balanced Loss Based on Effective Number of Samples”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 9268–9277. DOI: 10.1109/CVPR.2019.00949.

[38] M. Buda et al. “A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks”. In: *Neural Networks* 106 (2018), pp. 249–259. DOI: 10.1016/j.neunet.2018.07.011.

[39] C. Guo et al. “On Calibration of Modern Neural Networks”. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 1321–1330. URL: https://proceedings.mlr.press/v70/guo17a.html.