CIDAI Centre of Innovation for Data tech and Artificial Intelligence

T2 The computer vision centre (CVC) July 2022

Health medical image projects

Evolution of implementation and validation AI-bases models and CDSSs

Debora Gil Resina, Carles Sanchez

CIDAI-PAI 01-2021-D02

Suggested citation. CIDAI-PAI 01-2021-D02 // Health medical image projects - T2. Evolution of implementation and validation AI-bases models and CDSSs by Debora Gil Resina, Carles Sanchez, CIDAI, CVC 2021. Licence: CC BY-NC-SA 4.0

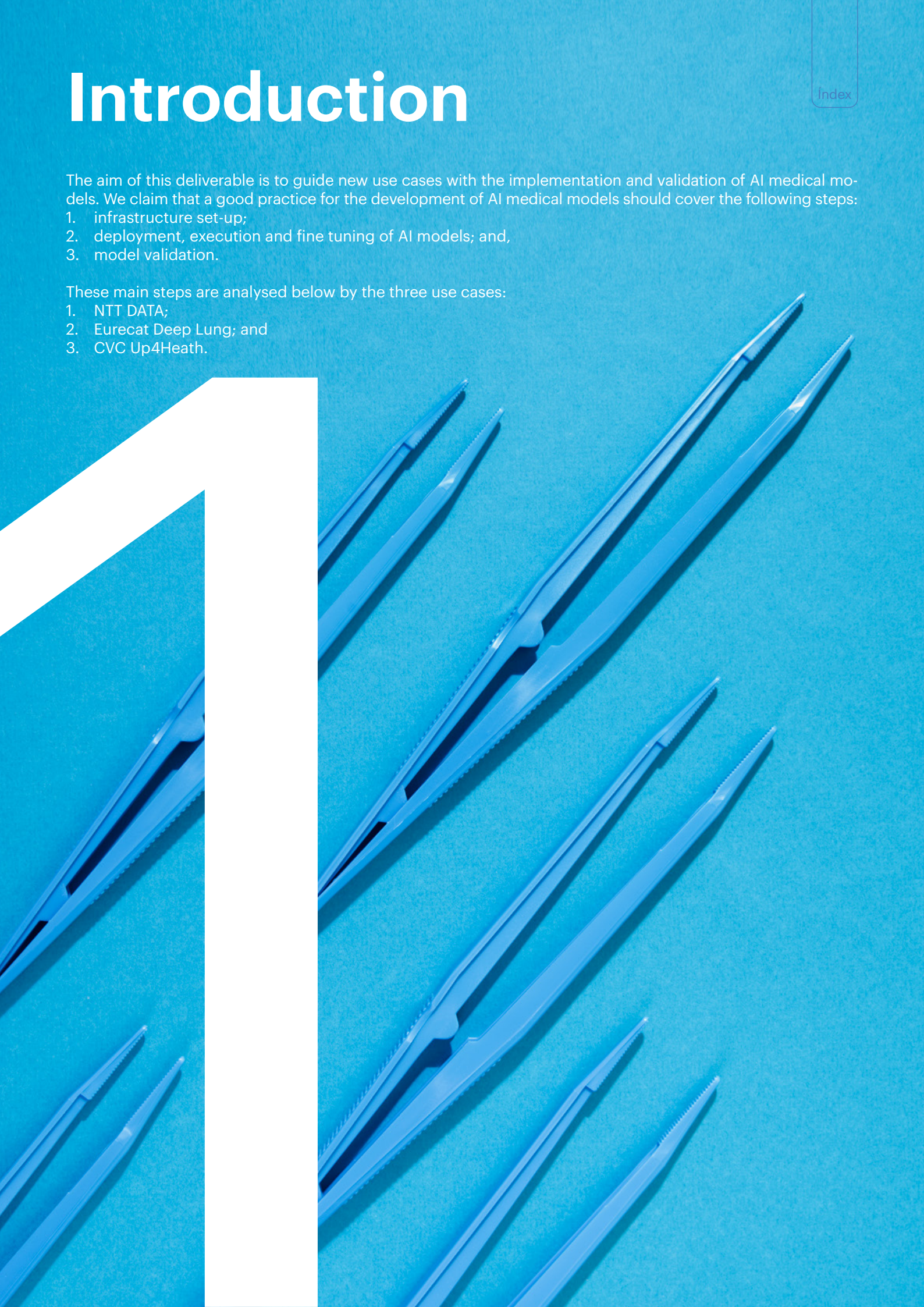
Introduction

The aim of this deliverable is to guide new use cases with the implementation and validation of AI medical models. We claim that a good practice for the development of AI medical models should cover the following steps:

1. infrastructure set-up;
2. deployment, execution and fine tuning of AI models; and,
3. model validation.

These main steps are analysed below by the three use cases:

1. NTT DATA;
2. Eurecat Deep Lung; and
3. CVC Up4Heath.



Infrastructure set-up

Distributed computing claims lower cost, high scalability, availability raised as a more attractive option to standalone solutions. Cloud-based services are being used in numerous areas such as education, engineering, healthcare, and more. In Healthcare, these services could deliver on-demand, self-service Internet infrastructure for medical imaging, entitled from healthcare practitioners to patients. One service that can be supported by cloud-based solutions will contribute and cover two main shortcomings: the performance improvement of research methods by gathering, and testing new data, and in the same way, to make research more accessible to the clinical community.

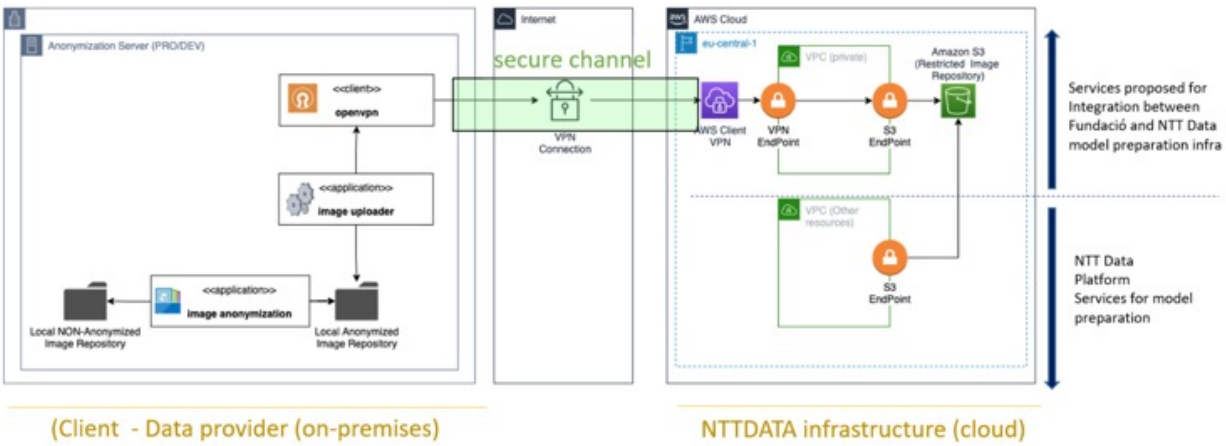
2.1. NTT DATA case

The AI project infrastructure has been deployed in a distributed way comprising two ends:

- One end, inside the data provider infrastructure it has been deployed the infrastructure (HW and software) to deploy the modules for data anonymization and data processing.
- The other end, it has been used cloud resources to deploy the AI model and required additional software and libraries to execute it as well as an application devoted to perform the medical image annotation, as this project follows a supervised learning approach.

The Figure 1. Scheme of the infrastructure. shows a scheme of the proposed infrastructure:

Figure 1. Scheme of the infrastructure.



On the left side, it can be observed the anonymization module – as a previous mandatory step before processing sensitive data) and the local repository which also includes the data processing engine to adequate the data (clean and homogenize the relevant data and metadata) to the required AI model data hierarchy.

The AI model together with required libraries and additional software has been deployed in the cloud to scale as needed. It includes also the services for the model preparation and validation.

2.2. Eurecat Deep Lung – Nodule Malignancy Classification

All AI developments in EURECAT take place at the DATURA platform. The aim of the platform is to provide an integrated platform with some of the best open source tools to handle all the needs of a data scientist and AI researcher. In Table 1 Data Storage Requirements.we present the different technologies used for the storage purposes.

Table 1 Data Storage Requirements.

Type	Technology	Description
Document Store	MongoDB	In many projects, data is stored in mongoDB for the analysis
Text Store	Elasticsearch	Many projects need full text search
Streaming store	Kafka	Kafka is one of an ideal solution for streaming applications
Columnar store	Cassandra	Columnar store to store structure data
Distributed storage	HDFS with Parquet	Data is stored as parquet files
Graph Store	OrientDB	For storing graph data
Key-value store	HBase	To store key-value data
Relational	MySQL / PostgreSQL	To store relational data

The software stack is summarized in Table 2 Software stack requirements.

Table 2 Software stack requirements.

Type	Technology	Description
Notebook	Jupyter, Zeepline	Python/Scala
Data science libraries	Numpy, pandas, matplotlib, scipy, scikit-learn, seaborn	Python
Distributed analysis	Spark (Pyspark)	Python/Scala
Dashboard	Kibana	-
Deep learning	Keras, tensorflow	Python/Java
Connectors for storage services	MongoDB, Elastisearch, Kafka, Cassandra, hdfs, OrientDB, Hbase, MySQL	Python/Scala/Java

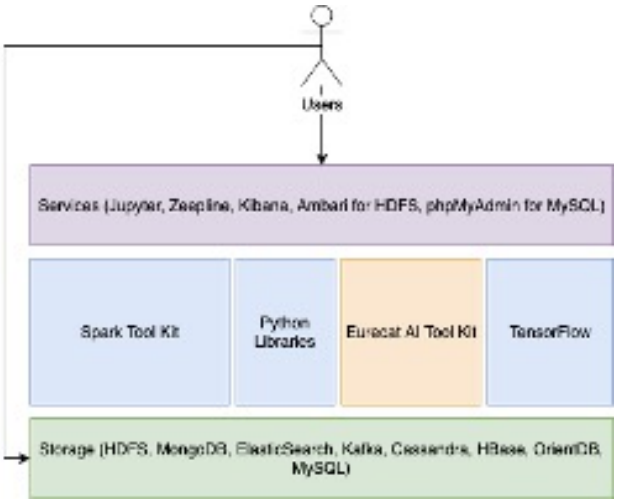
Storage Layer: It provides all required type of data stores for storing the data. It will also provide GUI for HDFS and MySQL. For others, user will be provided with the connection information and corresponding open source GUI to connect.

Processing Layer: Users can process their data on single or multiple machines. For single machine, user can use python to access their data for analysis. For multiple machines, users can use PySpark for analyzing their data. The connector for each store and libraries will be pre-installed.

Service Layer: This is the main layer exposed to users. Users will be provided with username and password to access the respective services. All the libraries will be installed to provide best user experience.

In Figure 2 Logical Architecture, we show the logical architecture of the data science platform as SaaS.

Figure 2 Logical Architecture



2.3. CVC Up4Health- Characterization of Lung Pathologies

In order to establish connection between new AI models and medical staff, a Client-server architecture for Diagnosis as a Service (DaaS) could be a robust solution [CVC1]. DaaS is a cloud computing model in which a provider hosts application in a server that clinicians use via internet. Since DaaS does not require to install applications on clinicians' own computers, it allows the use by multiple users of highly specialized software (implemented in different languages, as it happens in research) without extra expenses for hardware acquisition or licensing. A DaaS tailored for clinical needs not only would alleviate licensing costs, but also would facilitate easy and fast access to existing prototypes. The functional and computational requirements for the distributed computing architecture are the following.

Our front-end back-end will be controlled by an API service composed by different modules implementing specific services (sketched in Figure 3 DaaS architecture). Table 3 DaaS: Services, functional requirements and software.summarizes for each service its functional requirements and software. These modular architectures have been designed to allow for deployment in any sort of distributed system, edge, cloud or hybrid. The minimum hardware for a deployment of either a proof of concept prototype or an industrial functional system is detailed in Table 4 Hardware requirements for computation depending on the final system..

Figure 3 DaaS architecture

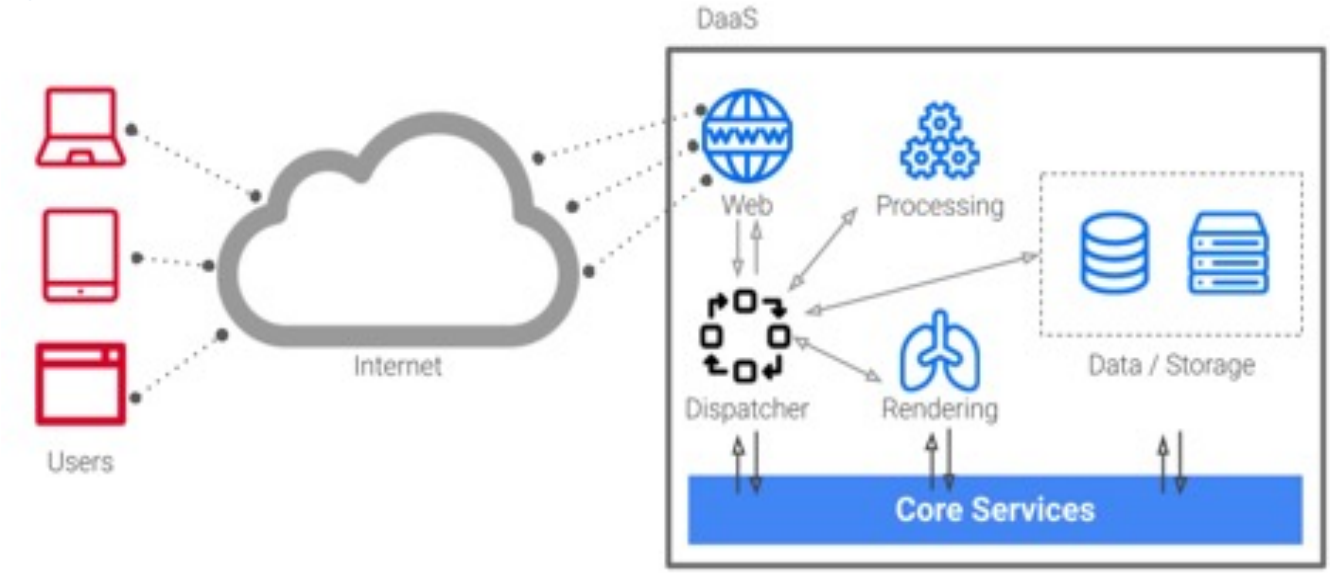


Table 3 DaaS: Services, functional requirements and software.

Service	Functional requirements	Software
Web	Manages the front-end	Webserver, php
Processing	Run (train/test) the models	GPU, Matlab, Python, c
Dispatcher	API controller	Python
Rendering	Shows the graphics	GPU, VTK, c
Core	Manages main functions	Python
Data Storage	Database system	Mongodb, mySQL

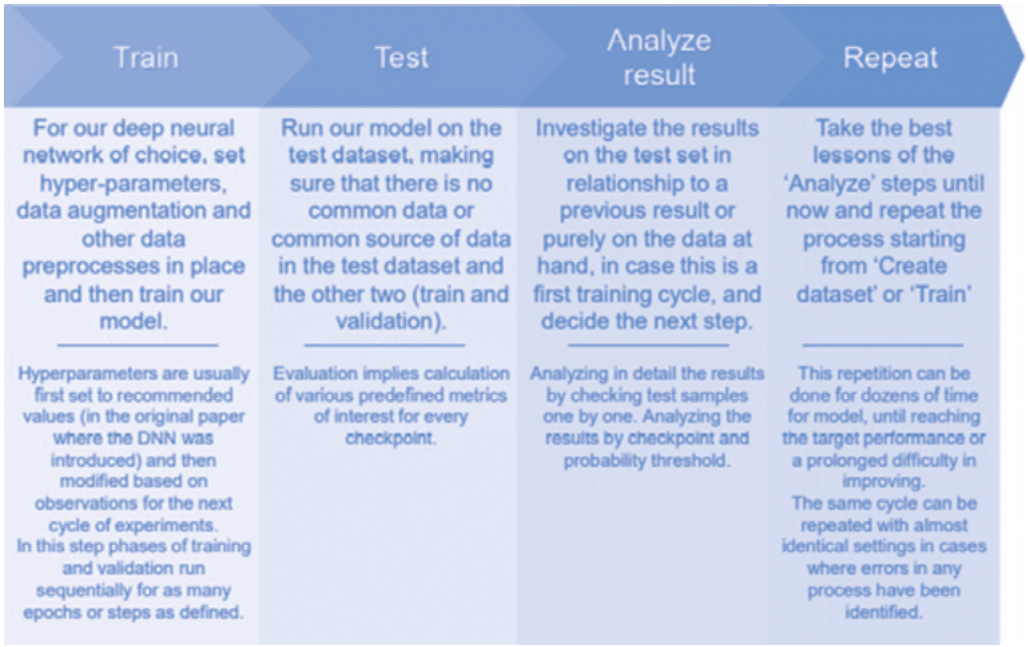
Table 4 Hardware requirements for computation depending on the final system.

Proof of concept	Functional system
CPU 2GHz with dual core	CPU 2.5GHz with quad core
4GB RAM	512GB RAM
3GB database/disk	5TB database/disk
	2GPU titanX for offline processing

Deployment and execution of AI models. Fine tuning and adjustments

3.1. NTT DATA case

The execution of the AI model has comprised the following steps:



Employed neural networks:

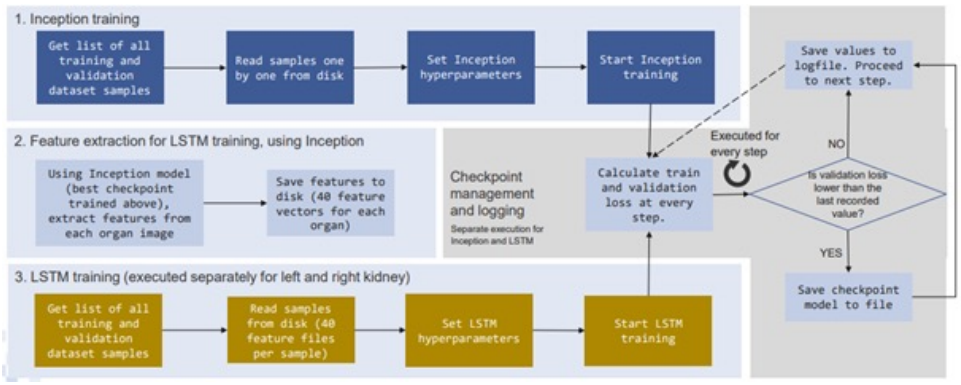
The following neural networks have been trained:

- Inception: trained to classify kidney images into normal and abnormal. By training Inception for this task, the neural network will actually learn to extract relevant features from kidney images regarding abnormalities. At the moment there are various versions of the Inception neural network used. The version used in this project is Inception V3. Inception V3 is a deep learning model based on convolutional neural networks (CNNs), which is used for image classification. Inception V3 is a superior version of the basic model Inception V1 which was introduced as GoogLeNet in 2014.
- LSTM: trained to classify entire kidneys into normal and abnormal. LSTMs (Long Short-Term Memory) are advanced RNNs (recurrent neural networks) that can handle long term dependencies. RNNs are a type of neural networks that have memory (which CNNs such as Inception do not have). RNNs work great on inputs that are sequences, as they are able to remember information from the past and use it to output the prediction of "what comes next". They can form a much deeper understanding of a sequence and its context compared to other algorithms. The traditional RNNs though have very "short memory", as in they can mostly only use the "recent information" to predict.

Monitored metrics for while training the models

When training the models, we keep track of different metrics in order to select the best performing epoch / step. The metrics are listed below:

- Train loss
- Sensitivity
- Precision N
- FI-score N
- Validation loss
- Specificity
- Precision A
- F1-score A



Training methodology

The training and validation methodology has comprised the following step-by-step process:

3.2 Eurecat Deep Lung

One of the first and most important tasks for radiologists in the lung cancer assessment is screening the entire lung CT volume, searching for small suspicious regions or nodules (usually between 3 mm to 30 mm) [EUT37]. Nowadays, this problem, as in most of the computer vision research areas, is being addressed through convolutional neural networks (CNN) [EUT8] able to extract, without human intervention, accurate feature image representations thanks to their shared-weights architecture and translation invariance characteristics. A common approach for automatic nodule detection consists of dividing the problem in two steps [EUT19,EUT20]: candidate detection and false positive reduction. In the first stage, 2D region proposal networks, such as faster region-based networks (Faster-RCNN) [EUT21], are used to extract suspicious regions of interest from the whole CT scan. In the second stage, these regions are classified on normal tissues or nodules using 3D CNN networks, in which the input are 3D image patches around the center of the nodules. Other recent approaches directly address this problem in a single step [EUT22-EUT24]. They re-adapt region proposal networks with 3D deeper architectures (such as ResNet [EUT35] or DenseNet [EUT36]) to directly predict 3D bounding boxes surrounding the nodules.

Another important task for lung cancer assessment is determining the size of the nodule. Currently, radiologists calculate the size of the nodule by visual inspection of the nodule on the CT scan, locating and measuring the largest diameter (in mm) [EUT38]. Usually, this measure is extrapolated to 3D dimensions, by means of mathematical operations [EUT25], to approximate the volume of the tumour. Although this process is simple and fast, it entails significant intra and inter variability in the size of the nodule, which it can reach from 2 to 3 mm in diameter [EUT26]. Since this variability may negatively impact the disease management, several deep learning solutions have addressed nodule size measurement to support clinicians. Some works [EUT22,23] propose learning the diameter of the tumour by extending the nodule detection network (either in 2D or 3D) with a new output in the network. Other solutions build semantic segmentation networks to automatically determine the pixels of the nodules with which later extract the nodule diameter or volume size. One of the most common and successful architectures for segmentation is the U-Net [EUT27]. This type of networks uses a convolutional encoder and decoder backbone, tied at different levels by short-cuts, which allow to by-pass high level features of the encoder to the decoder, to enhance the image reconstruction task. Several extensions of this architecture can be found, such as in [EUT28] building a 3D version of it, or in [EUT29] incorporating ResNet-like blocks and a Dice-based loss layer, more suitable for segmentation tasks. A more recent approach, nnu-net [EUT30] has been successfully applied to a multitude of medical segmentation problems (including pulmonary nodule segmentation). One of the benefits of this approach is the automatic fine-tuning of several configuration parameters to the images to be segmented.

Despite the high performances reported by U-Net like networks, they address the segmentation problem from a deterministic point of view. However, due to the inherent ambiguity of the problem (often contours of the nodules are not clearly delimited), it is desirable reporting network uncertainty estimates when predicting the size of the nodules. One way to learning model uncertainty is moving from networks that models the problem from a 1 input to 1 output solution to modelling the problem to 1 input to N possible valid outputs. This change of paradigm has already been tackled in deep neural networks through different approaches. One of the simplest approximations consist of ensemble multiple networks to provide multiple opinions [EUT32]. Another approach consists of enabling dropout [EUT33] at inference time to provide independent pixel-wise probabilities [EUT34]. Other approximation is through deep generative networks. One of the most well-known examples are the generative adversarial networks [EUT59]. This type of networks tries to learn, in an unsupervised manner, a direct mapping from a random noise vector to an output image. To do this, a generator network creates new valid images (from random vectors) with the intention to fool a discriminator network that evaluates whether an image is valid or fake. An extension of this type of networks are conditional GANs [EUT39], in which the goal is to learn structured outputs conditioned on an input image. To do this, the discriminator receives as input the target image to which conditioning the generator. Like cGANs, we can find the conditional variational autoencoders (CVAE) [EUT34]. This type of networks proposes learning a multi-dimensional latent space that encodes all possible output images. During training, the latent space distribution defined by the encoder is

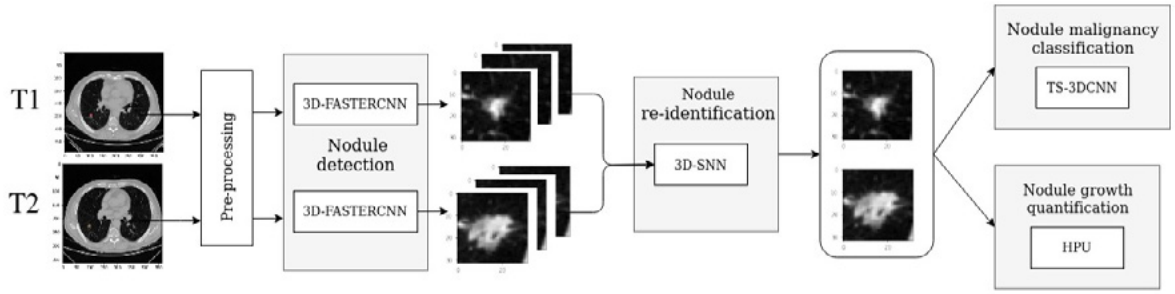
approximated to a normal distribution to assure continuity and avoid 'mode collapse' commonly seen in GAN approaches [EUT60]. Also, the random vector sampled from the latent space, together with the target image are passed to the decoder (only during training) to generate a new plausible image. A recent work, hierarchical probabilistic U-Net (HPU) [EUT40], has been proposed to cover the gap between the generative ability of producing new structured images of the CVAE, with the accuracy of segmenting images of the U-Net. To do this, during training, a posterior U-Net like network, conditioned on the radiologist ground truth nodule, is added to transfer the latent features to a prior U-Net like network by injection, at different levels of the decoder part of this network.

Another crucial task for radiologists is the lung cancer diagnosis. To support them in this task, several works have been proposed relying on 2D and 3D inputs, using different deep learning architectures (e.g. CNN, RNN) [EUT41-43], but mostly relying on single CT scan images (commonly derived from the LIDC dataset [EUT49]). Therefore, very few deep learning works have addressed the temporal evolution of pulmonary nodules to support the clinical decision-making. In [EUT44], an end-to-end deep learning-based pipeline was presented for lung cancer prediction using two CT studies per patient (current and previous year). This approach proposes three 3D CNN networks, one for analysing the image at lung CT level, other at nodule patches level, and a final one, to provide cancer risk prediction using outcomes from previous two components. In [EUT52], a deep learning approach is proposed for lung cancer risk at 3 years and lung cancer-specific mortality. In this study, although it is not focused on automatic image analysis, they use a multilayer perceptron to ensemble nodule and non-nodule features associated to lung abnormalities. Like in [EUT44], they use the data from the National Lung Screening Trial (NLST), but results were validated in another longitudinal study [EUT53]. In [EUT17], the authors built an automatic pipeline for lung nodule growth quantification using data from a longitudinal cohort of incidental nodules. In this approach, instead of applying lung CT image registration [EUT58], the authors propose a 3D Siamese CNN to re-identify nodules from different CT studies of the patient. That pipeline also allows nodule growth assessment using the diameters predicted by the nodule detection network.

Temporal Lung Nodule Assessment

Our pipeline takes as input two images from the same patient at different time-points, identifies the lung nodules, and estimates their malignancy and growth. The pipeline consists of 4 main components (see Figure 1): 1) nodule detection, which is done independently on each image; 2) nodule re-identification, which finds the correspondence between nodules across time points; 3) nodule malignancy classification; and 4) nodule growth quantification.

Figure 1. Pipeline and architecture for the temporal analysis of lung nodules



The following subsections describe each component, making special emphasis to which are their input and outputs, the core methods, and their configuration.

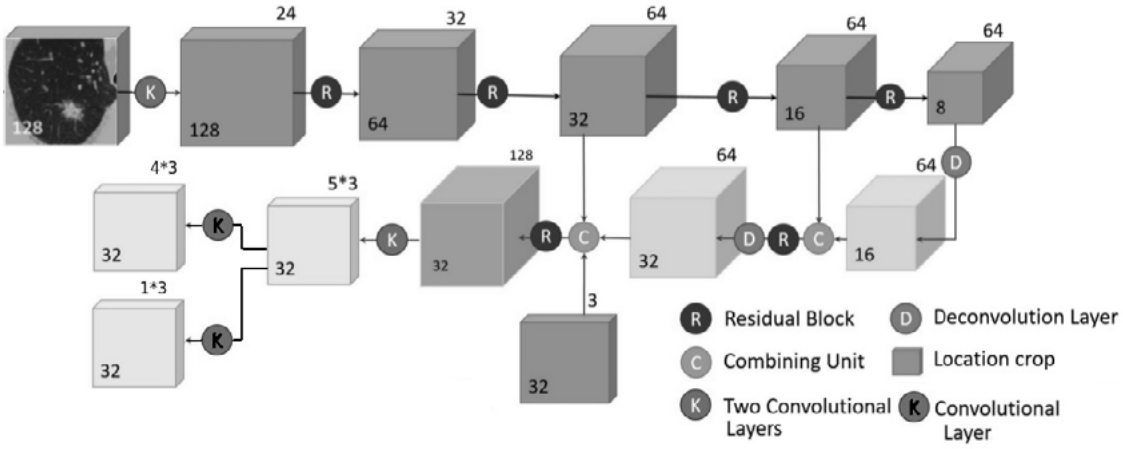
Pre-processing

Lung CT images are usually originated by different scanners and at different image resolutions. Therefore, the pipeline makes an initial pre-processing step with the intention to standardize the input images. Precisely, first, the images are resampled to an isotropic resolution of 1x1x1mm. Second, the image pixel intensities are clipped between [-1000, 600] Hounsfield Units to filter out non-tissue related regions. Finally, the pixels are normalized between 0 and 1.

Nodule detection

The first component of the pipeline consists of detecting pulmonary nodules in CT scan images. To this end, we adopted the solution presented in [17] which relayed on two previous and successful works [EUT22,EUT23]. The neural network follows a Faster-RCNN scheme [EUT21] adapted for 3D images of [128x128x128]. The backbone of the network was like the U-Net [EUT27] architecture. The output of the network was the location of the nodules (x, y, z coordinates), the diameter, and a probability of being nodule. The loss function was the sum of mean squared errors, for each of the nodule location and diameter parameters, and binary cross entropy for the probability output. The training was configured with a batch size of 8, Adam as optimization algorithm, and a learning rate of 0.1 with a decay of 0.001 every 100 epochs, with a total of 450 epochs. Online hard example mining [EUT47] with a factor of 20 times the batch size was used to optimize the network. To reduce overfitting, 3D data augmentation (such as random rotation, flip, and zoom in/out) was used during training. The network architecture is shown in Figure-2.

Figure-2. Architecture proposed for the nodule detection network of the pipeline.

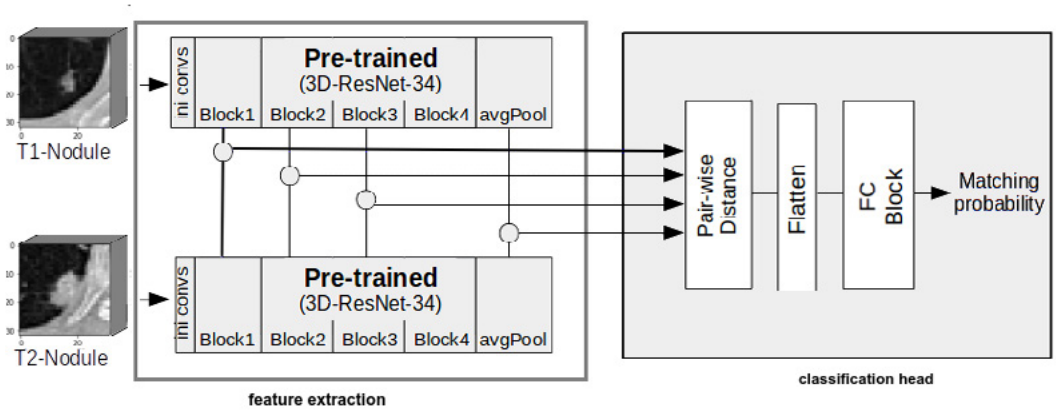


Nodule re-identification

Once the nodules from two different CT scans of the same patient are detected, a second component aims to match or re-identify these nodules. To automatically achieve this task, a 3D siamese neural network (3D-SNN) presented in [EUT17] was adopted. A SNN [EUT48] is made up of two components: feature extraction and classification. In the first, two subnetworks (with shared architecture and weights) process a pair of images at a time to produce two embedding feature vectors directly from the images. In the second, a head network determines whether the two embedding feature arrays are similar (i.e., correspond to the same nodule).

From the different re-identification network setups presented in [EUT17], we used the one that obtained the best results (FIFB). This setup consisted of freezing the sibling networks of the feature extraction component with the weights of a pre-trained network, initially built for nodule identification [EUT16]. According to [EUT17], from the different convolution blocks of this pre-trained network, we used the output from the first block, as inputs for the head component of the 3D-SNN, since they reported better performances. The classification head component was configured with a L1-pairwise distance, a flattening layer, and a fully connected (FC) block, comprising a FC layer (with 64 units), a batch norm, a ReLU, a dropout layer and a final FC layer (with one unit). The 3D-SNNs was trained using binary cross-entropy loss function, 150 epochs, 1e-4 of learning rate, 8 of batch size, 0.3 of dropout, 10 epochs for early stopping, and Adam as the optimization algorithm. Moreover, random rotation, flip, and zoom were applied for data augmentation. Figure-3 shows the SNN architecture for the nodule re-identification problem.

Figure 3. Architecture of the 3D-SNN for the nodule re-identification component of the pipeline.



Nodule growth quantification

A couple of methods were proposed for the nodule growth quantification component. The first one, already used in [EUT17], consisted of computing the diameter difference of the paired nodules from the re-identification component. The diameters of both nodules were taken from the nodule detection component of the pipeline.

The second approach consisted of using a hierarchical probabilistic Unet [EUT40]. This type of networks can generate unlimited segmentations given an input nodule image. To do this, two networks the posterior and the prior work collaboratively to segment the nodules. In particular, the posterior network learns to model the conditional joint probability distribution between the nodule Ti at a time-point i and its segmentation (ground truth) Si. Meanwhile, the prior network tries to model this same distribution but only having as input the nodule Ti.

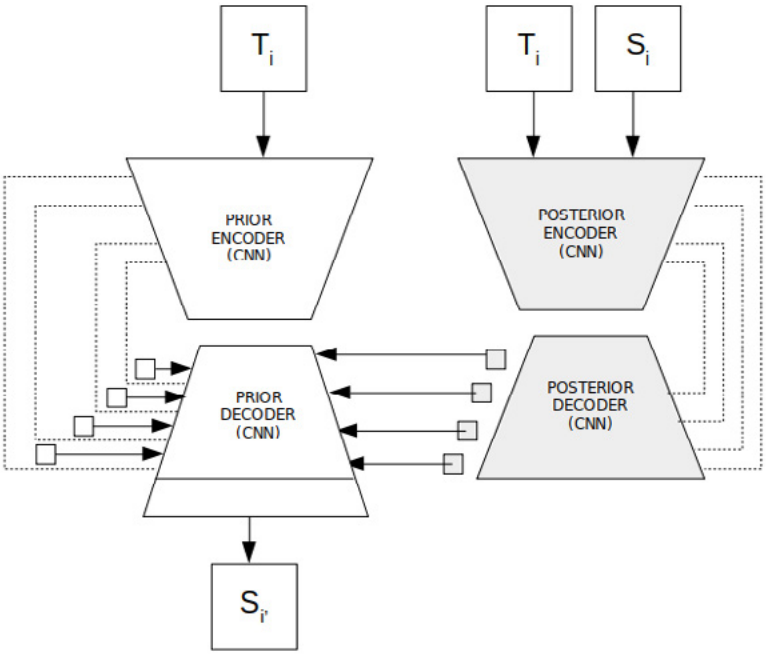
During training, both networks are tied together to allow gradient propagation with a stochastic gradient descent-based mechanism. The way to do this, is by injecting the different latent features, defined at different levels of the decoder of the posterior network, to the corresponding following layer but from the decoder of the prior network.

The loss function to be minimized, or elbo function, is composed by the distance between the prior and posterior distributions (kullback-leibler divergence), and the quality of the resulting generated segmentation (cross entropy loss).

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{\mathbf{z}\sim Q}\big[-\log P_c(Y|S(X,\mathbf{z}))\big] + \beta \cdot \sum_{i=0}^{L} \mathbb{E}_{\mathbf{z}_{<i}\sim Q} D_{\text{KL}}\big(q_i(\mathbf{z}_i|\mathbf{z}_{<i},X,Y)||p_i(\mathbf{z}_i|\mathbf{z}_{<i},X)\big).$$

At inference time, a random sample from the distribution defined by the different latent features, located at different levels of the decoder of the prior network, are injected into the following layer of this same network to output a new nodule segmentation.

Figure-4. Hierarchical probabilistic Unet network architecture overview. Ni is the nodule image i, Si is the ground truth segmentation of the nodule Ni, and Si' the predicted segmentation of the nodule Ni



The authors [EUT40] evaluated this network for the lung segmentation problem using data from the LIDC dataset [EUT49]. To do this, they employed the generalized energy distance, a metric to account for quality of the segmentation and variability in generating segmentations, according to the variability in the ground truths.

$$D^2_{\mathrm{GED}}(P_{\mathrm{gt}}, P_{\mathrm{out}}) = 2\mathbb{E}\Big[d(S, Y)\Big] - \mathbb{E}\Big[d(S, S')\Big] - \mathbb{E}\Big[d(Y, Y')\Big],$$

In the original paper, the HPU network was trained with labelled segmentations from up to four radiologists. When no nodule was marked by a radiologist an empty segmentations image was used. This made the network to model as well the probability of not finding a nodule to segment in an axial CT slice. However, for our settings, this was undesired as the nodule detector already filter out non-nodule cases. Thus, we retrained the HPU network using the same specifications and architecture as in the original paper but omitting empty segmentation cases.

Then, we used the HPU (prior) network to estimate the nodule growth and a measure of dispersion (standard deviation). To do this, we ran N times (N=1000) the HPU network for the nodule T1, obtaining N segmentations. From these segmentations we derived the major diameter nodule, obtaining a random vector of N diameters. We repeated this same process but for the nodule at T2. Then, assuming Gaussianity for simplicity, we obtained the diameter growth mean as the mean difference of both random diameter vectors, and the standard deviation, as the squared root of the sum of the variances divided by the size of the samples.

Nodule malignancy classification

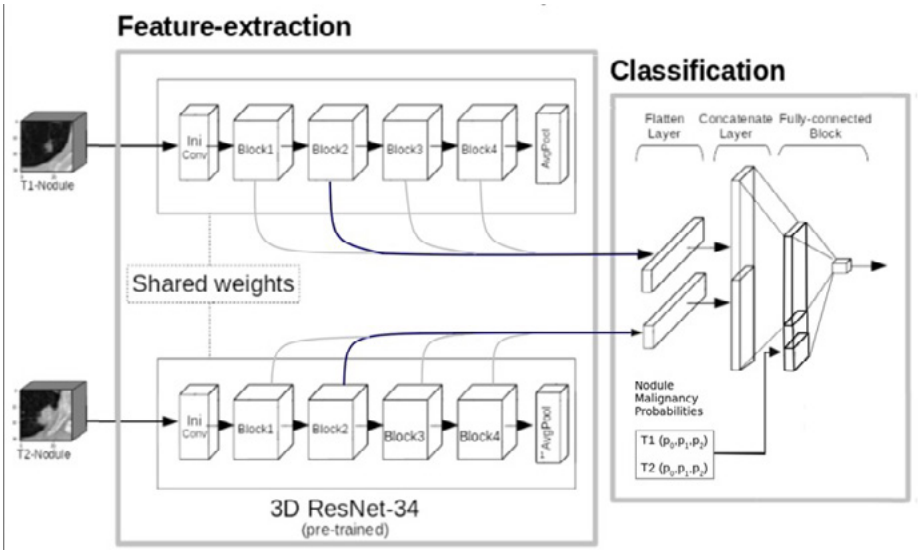
The problem of nodule malignancy classification was addressed with three different approaches, one using nodule malignancies annotated by radiologist (i.e. not cancer confirmed cases) from a single time-point nodule image, and the other two using confirmed diagnosis nodule malignancies (either from biopsy or without significant growth increase during at least 2 years) with two patches of the same nodule taken at two different time-points.

The first approach consisted of re-using the best nodule malignancy network extracted from [EUT16]. This network (3D-CNN-MAL) receives as input a single volumetric nodule of 32x32x32. The network has a tailored architecture composed of 4 blocks of 3D CNNs, interleaved with dropout and a final dense layer with a softmax layer at the end. The outputs of this network are 3 probabilities corresponding to 3 categories of nodule malignancy (benign, suspicious, and malignant).

The second approach was extracted from [EUT18]. As in the previous approach, we selected the model with the best performance (TS-3DCNN). This model expected two volumetric input patch images of 32x32x32 centered around the nodule. The two patches corresponded to the two CT scans made on the same patient but at different time-points. The network was a two-stream 3D convolutional neural network, in which two feature extraction sub-networks, with same architecture and weights, analyzed in parallel the nodule patches, while the classification network part provides a cancer probability risk. Due to the difficulty of collecting enough longitudinal lung cancer CT images for training deep neural networks, the siblings of the TS-3DCNN were transferred from a pre-trained 3D ResNet-34 network, used for identifying pulmonary nodules [EUT16]. We used the features from the last layer of the second block of the 3D ResNet-34, as the ones which reported better performances. The classification head component of the TS-3DCNN was configured with a flattening, a concatenation, and a FC block layer comprising a FC layer (with 64 units), a batch norm, a ReLU, a dropout and a final FC layer (with one unit). Figure-5 shows the architecture of the TS-3DCNN network.

Inspired by the promising results of [EUT16], we proposed an extension of previous second approach (TS-3DCNN-MAL). We proposed to, given a new nodule to classify, predict the malignancy probability, using the previously described 3D-CNN-MAL network, and then integrate them in the TS-3DCNN network. Precisely, 6 extra features (corresponding to the 3 outcomes of the 3D-CNN-MAL for each time-point nodule) were concatenated with the features of the last fully connected layer of the TS-3DCNN.

Figure-5. Nodule malignancy classification architecture of the TS-3DCNN-MAL network.



3.3 CVC Up4Health- Characterization of Lung Pathologies

A main challenge in the application of deep learning to biomedical problems is the limited amount of good quality data with annotations, which is a must for training new models with complex architectures. Besides, in the case of benign nodule screening,

this is aggravated with the fact that the problem is highly unbalanced with benign cases being the minority class. Under such experimental settings, models are often over-fitted [CVC2] results are non-reproducible [CVC2,CVC3] and most times [CVC4,CVC5,CVC6,CVC7] do not outperform conventional machine learning approaches [CVC8]. Another pitfall, especially for deep methods is models should also be easily interpreted from a clinical point of view to allow the analysis of the clinical factors that have an impact on the clinical decision [CVC9].

To address the above challenges, we propose hybrid approaches that embed lesions into a low dimensional radiomic space defined by classic radiomic features. These features are the input to a fully connected classification network with an architecture optimized to ensure maximum clinical outcome. Our intelligent radiomic system has three main steps. First, scans are pre-processed. Second, radiomic features selected according to its reproducibility are extracted from the registered volumes to define a radiomic feature space. Finally, a machine learning method is used to disseminate each value of the feature space between the different types of lesion patterns.

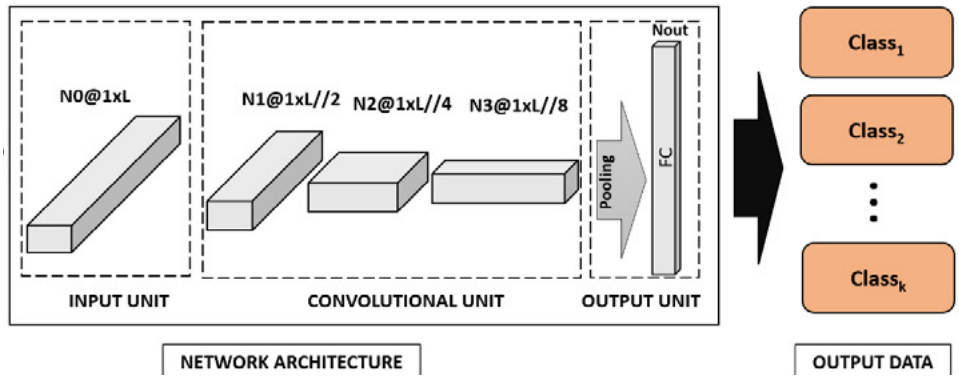
The pre-processing step depends on the input scans specific to each pathology: CT- SPECT for pulmonary embolism, CT for nodule diagnosis.

The radiomic features are a subset of PyRadiomics, an open-source python package for the extraction of Radiomics features from medical imaging volumes. PyRadiomics features include shape features, first order features, and textural features (Gray Level Co-ocurrence Matrix (GLCM), Gray Level Size Zone (GLSZM), Gray Level Run Length Matrix (GLRLM)and Gray Level Dependency Matrix (GLDM)) describing several aspects of the lesion. The subset can be selected according to either the reproducibility against different image acquisition conditions and interobserver variability in lesion identification [CVC10, CVC13] or correlation to malignancy [CVC11].

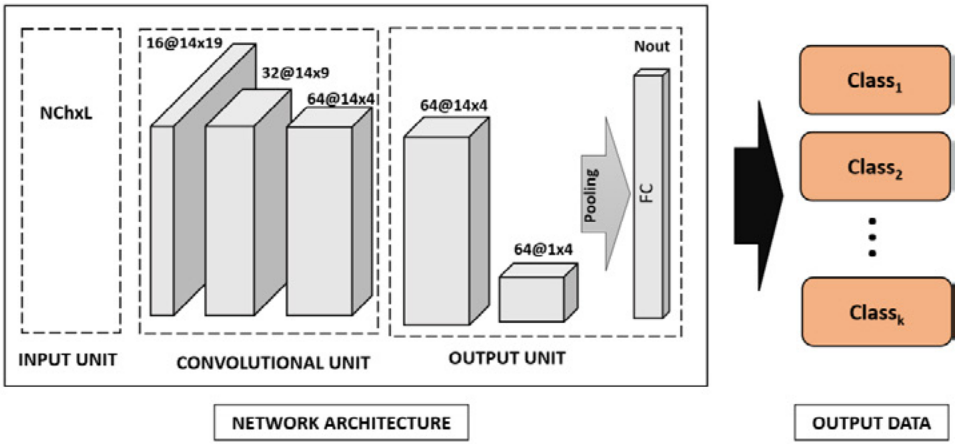
These radiomic features are input to fully connected networks with different strategies for the combination of multi-modal data [CVC12] at two different levels: input data (input projector model) and convolutional feature (feature projector model) models. Both approaches consist of an input unit managing fusion at the input level, a convolutional unit, and an output unit for fusion of convolutional features. Figure 4 shows the general architecture of each approach. For each approach, we optimized: 1) the number of layers; 2) neurons per layer, 3) optimizer; 4) learning rate; 5) weight initialization and ; 6) epochs.

Figure 4. Architectures for multi-modal data fusion: input projector model (a), feature projector model, (b).

Input Projector Model



Feature Projector Model



Architectures (number of layers and neurons for each layer) were optimized using the Optuna [CVC14] hyper-parameter optimization framework and a double k-fold stratification. First, a set of 20% of lesions is randomly selected to be held out as an independent test set. Second, with the remaining data, we use a nested cross-validation (CV) approach that is composed by an outer CV and an inner CV, used to make model selection and hyper-parameters tuning respectively [CVC15].

3.3.1 Detection of Pulmonary Embolism in COVID19 Pneumonia

In this case, since the final diagnostic requires the combination of information from the SPECT and CT scans, the pre-processing step consists in the registration of the two volumes in order to fuse both image modalities. In order to account for differences in SPECT intensities due to variations in contrast agent concentrations, SPECT scans were normalized using two different approaches. The first approach normalized SPECT values in the range [0, 1]. SPECT might have spots of high concentration of the agent that are not associated to a higher lung perfusion. These artefacts associated to the technique itself, deviate the maximum intensity values and introduce a drop in the intensity of normalized volumes not related to a drop-in perfusion. In order to alleviate the impact of this artefact in the predictions of pulmonary embolism (PE), we have also considered a normalization based on the superior quantile of SPECT intensity values.

This use case uses a selection of features based on reproducibility.

The proposed radiomic system analyses the intensity values of CT and SPECT volumes for each voxel in order to discriminate three clinically relevant types of image patterns defined by pulmonologist and nuclear medicine experts:

1. Control: healthy - normal CT and SPECT.
2. Pulmonary Embolism (PE): normal CT with a localized perfusion defect in the SPECT.
3. Pneumonia: affected CT but normal per-fusion in the area or with an atypical perfusion defect that does not meet PE criteria

$$Pixel_{value} = \frac{\frac{\text{HU} - \text{Intercept}}{\text{Slope}}}{\max(HU) - \min(HU)} * MaxIntensity$$

$$Pixel_{value} = \frac{\frac{\text{HU} - \text{Intercept}}{\text{Slope}}}{4000} * MaxIntensity$$

In order to avoid over fitting, two extra categories were added:

1. Black Background: areas with no tissue uptake.
2. Body Tissue: areas of the body not belonging to the lungs.

Further details about this use case can be found in [CVC10, CVC13].

3.3.2 Diagnosis of Nodule Malignancy for Lung Cancer Screening

In the preprocessing phase we used the anonymized CT-chest DICOMs and the annotated nodule ROIs. A nodule ROI always includes the intranodule region (inside nodule region), but depending on the nodule shape, the perinodular region (around nodule region) is included in greater or lesser extent. Since, in [CVC16] it is reported that the importance of using perinodular region in the benign and malignant nodules classification, we enlarged the ROI size a 15\%.

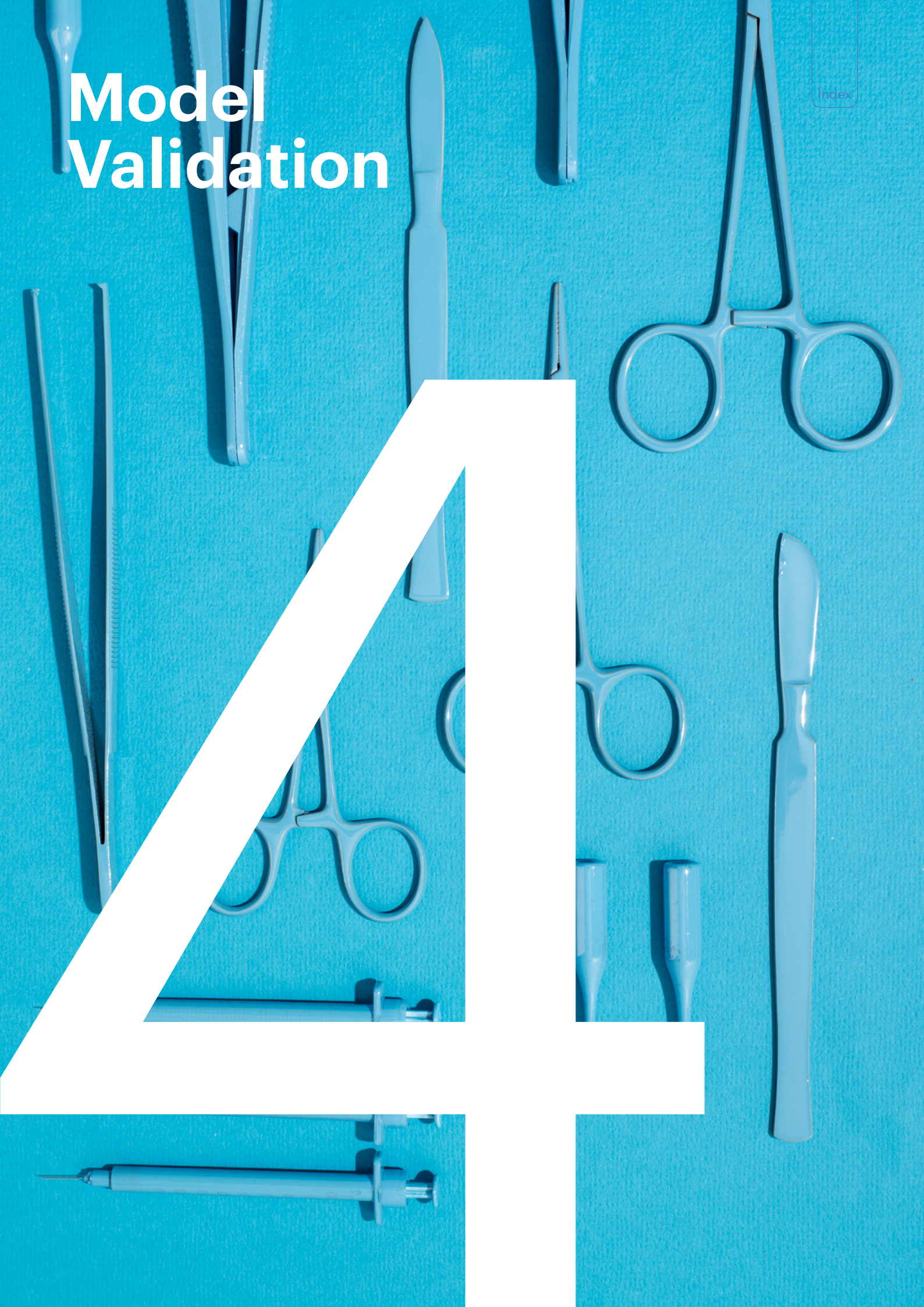
In order to segment the nodule, we applied Otsu thresholding to the ROI volume. Since the segmentation of peripheral nodules can include non-pulmonary tissue, the binarized volumes are masked with a segmentation of lungs. The final nodule segmentation was the largest connected component of the masked volumes. The segmentation of lungs was computed using thresholding and morphological operations [CVC17]. CT lungs were selected as the larger connected component of the voxels with intensity between 950 to -300 Hounsfield Units, followed by a closing with a structuring element of size 5.

Hounsfield Units were mapped to the common intensity range, [0,MaxIntensity] using the following formula to compute Pixel_value:

where max(HU) = 2976 was computed as the maximum of the Hounsfield Units of the training set volumes and min(HU) = −1024 is the HU value for air. The intensity range MaxIntensity and number of bins, Nbins [30, 130], are hyper-parameters that were set using grid-search. The optimal values were MaxIntensity = 24 and Nbins = 128.

For this use case we used a selection of GLCM features based on correlation to clinical output. Further details about this use case can be found in [CVC11].

Model Validation



4

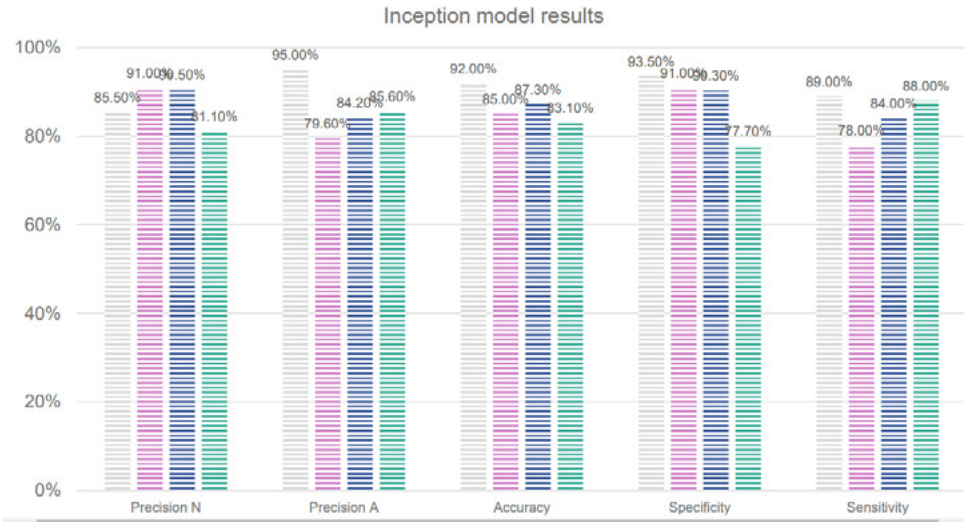
4.1. NTT DATA case

NTT DATA model is based on a set of complex neural networks and image processing techniques in the scope of judging whether an organ is normal or abnormal.

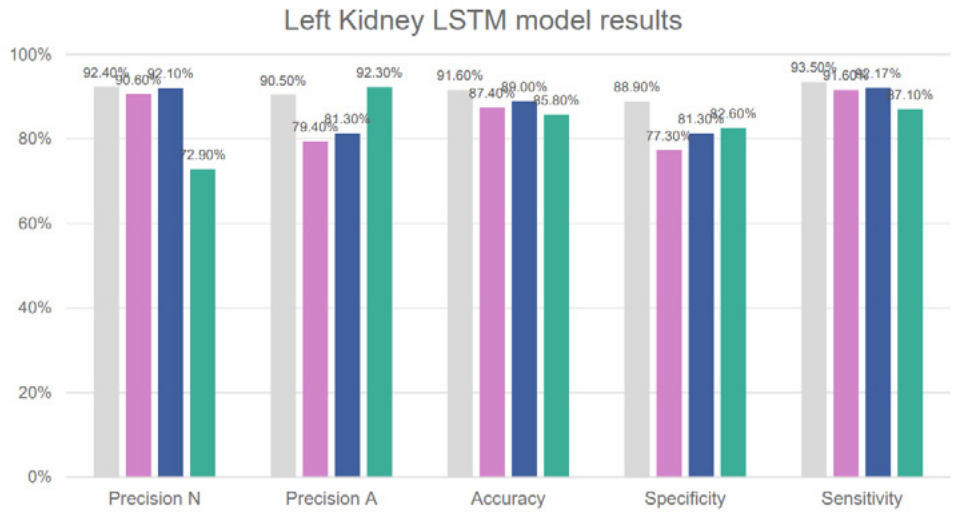
To this end, we have trained 3 different sets of models for Maestro Kidney Classification: Inception, LSTM for Left Kidney (LSTM LK) and LSTM for Right Kidney (LSTM RK). We measured and present in this report the following metrics for the best Inception, LSTM LK and LSTM RK models we have trained:

- Sensitivity
- Precision N (Precision of the Normal class)
- Precision A (Precision of the Abnormal class)

4.1.1 Results summary for Inception models

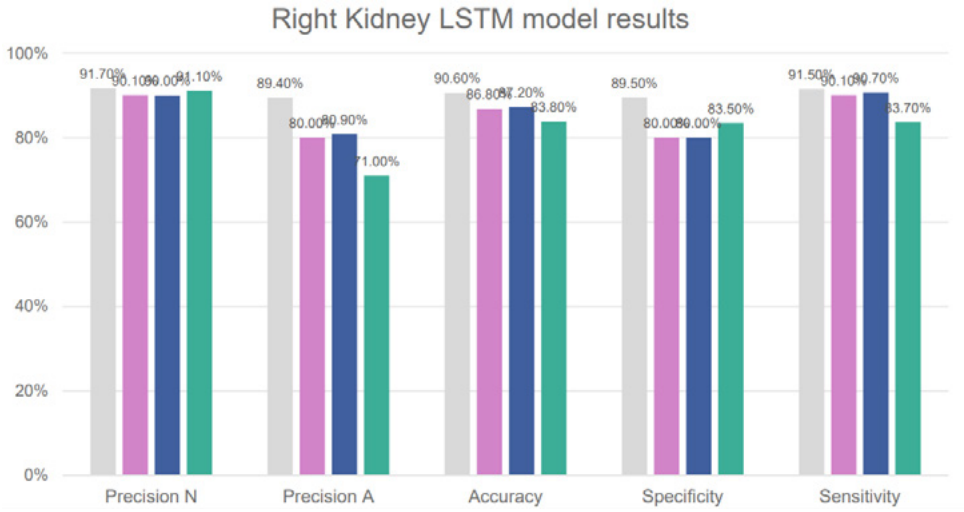


4.1.2 Results summary for LSTM – Left Kidney



By comparing the results of all LSTM LK models trained, we can observe that the best overall model on the data provided by the data provider organization is US2-C (blue), when the US2 model was fine tuned for this data.

4.1.3 Results summary for LSTM – Right Kidney



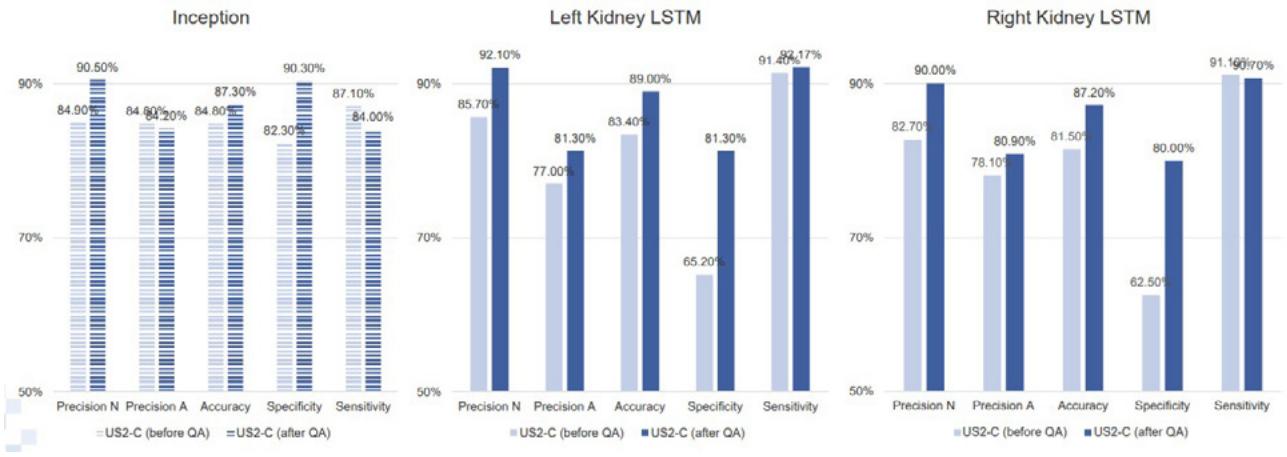
For RK the US2-C model (blue) is a clear option regardless of preference, because in this case it outperforms other models in most metrics, with minimal differences in case of Precision N and Specificity.

The biggest difference with the results of LSTM LK comes for Precision A, the C model performing extremely poorly in comparison to US2-C and its LSTM LK counterpart. We can suppose this is because of the presence of the liver and other clinical differences between left and right kidney, rather than an issue within the training of the AI.

4.1.4 Impact of annotation QA on AI training

One of the tasks conducted during the project has been to carry out a QA process on the annotations performed by the AI model. This QA process has been addressed by specialist radiologists on a set of studies that have been identified specifically to this end.

The charts below show the performance of the best performing models (models US2-C) trained on the annotations delivered (left side bars, faded blue) and models trained on the annotations delivered after QA by the team of radiologists' annotators was executed (right side bars, stronger blue).



A general improvement of all models can be observed, with the biggest difference being observed for Specificity. Since Specificity measures the ability of the models to identify the absence of abnormalities, we can assume that by correcting the annotations, the model's ability to identify feature of normal slices and normal kidneys has improved. Thus, it is important to highlight that the validation of AI based on the expertise of radiologists (or clinical experts in the scope of the speciality we are dealing with in general) is fundamental to improve AI performance in the scope of Health.

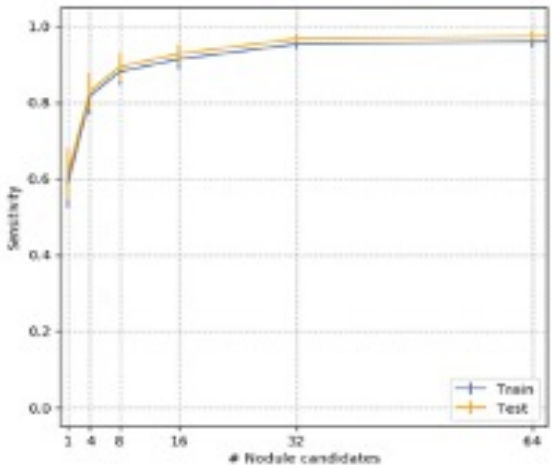
4.2. Eurecat Deep Lung – Nodule Malignancy Classification

Nodule detection

First of all, we evaluate the ability of the pipeline to detect the annotated nodules (one per each CT) among all nodules predicted by the 3D-FasterCNN network. Due to the VHLung cohort does not have annotated all possible nodules for each CT, but the most relevant for the clinicians, we measured the performance of the network to find the most relevant one in the least number of predicted nodules. Hence, the fewer number of predicted nodules to find the selected nodule, the better. To do this, we proposed different thresholds (1, 4, 8, 16, 32, and 64) or number of predicted nodule candidates, and we computed per each CT (of the test set) whether the annotated nodule was in each subset of predicted nodule candidates (ranked by probability). Taking a reasonable threshold of 32 predicted nodules, the pipeline on the 76 CT scans of the test set (taken individually) only in 2 of them the annotated nodule was not found among the candidates. To have a better estimation of the nodule detection performance, we repeated this process on 10 subsets, sampled randomly with replacement. Results showed a sensitivity of 0.973 for localizing the relevant nodule on 32 predicted nodules. This represents a missing ratio of 2.5+/-1.02 nodules in 76 CTs. Results can be shown in the FROC-curve

of Figure 5 FROC-curve of the malignant nodule detection algorithm for training and testing partition..

Figure 5 FROC-curve of the malignant nodule detection algorithm for training and testing partition.



Nodule re-identification

To evaluate the second stage of the pipeline, we used the 36 pairs of CT scans where the pipeline found the radiologist's annotated nodules. For each pair of CTs, we input 64 (32 per each time-point) nodule candidates into the 3D-SNN network to obtain those that correspond. In total, the 3D-SNN network reported only 4 CT-pairs incorrectly matched with an accuracy of 0.888. Table 5. Performance of the re-identification component of the pipeline. provides a summary of the results for the re-identification step, stratifying them by the initial size of the nodules.

Table 5. Performance of the re-identification component of the pipeline.

	Small	Medium	Large	Total
Accuracy	1.0	0.84	0.75	0.888
#Nodule-pairs	13	19	4	36

Nodule growth quantification

Two different approaches for nodule growth quantification were evaluated on the 32 matching nodules obtained from the nodule re-identification component. The first approach used the predicted nodule diameter measurements from the 3D-FasterCNN network, while the second used the predicted nodule diameter measurements from the HPU network. For a proper usage of the HPU in the context of the nodule growth quantification, we retrain this network according to [EUT40] with same data from LIDC dataset but omitting "empty" cases where radiologists did not mark any nodule in the axial slices. The model reported a GED2 of 0.38 and a reconstruction Dice of 0.91. Table 6. Nodule growth performance comparison between FasterCNN and HPU networks. shows the mean absolute error, mean squared error and r-coefficient of correlation respect to the ground truth. Results show the mean and 2 standard error associated with a 95% of confidence, obtained with a 1000 bootstraps with replacement of the test set. Figure 6. Comparative of radiologist growth measurements with results from nodule detector and nodule segmentation. shows estimated nodule growth sizes from both networks per each nodule of the test set.

Table 6. Nodule growth performance comparison between FasterCNN and HPU networks.

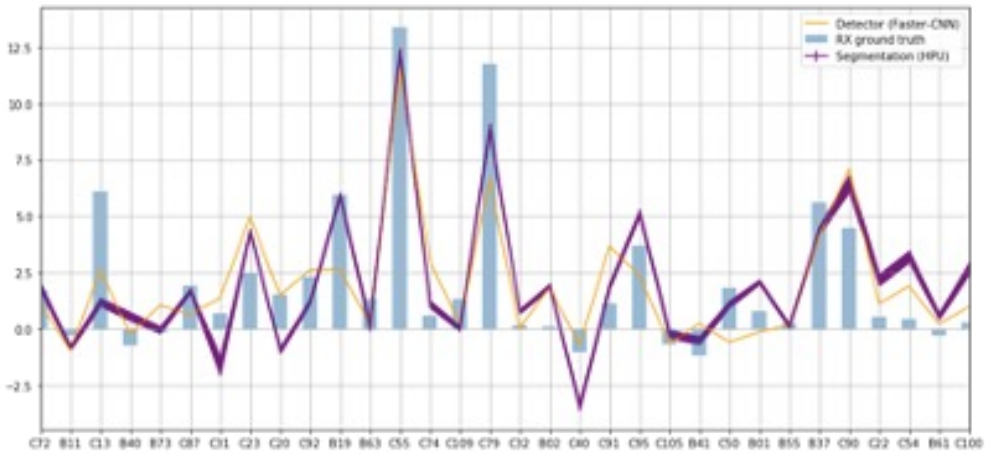
	Small	Medium	Large
3D-FasterCNN	1.400 +/- 0.422	3.333 +/- 1.927	0.637 +/- 0.344
HPU	1.348 +/- 0.370	2.889 +/- 1.561	0.667 +/- 0.418

These results show that the HPU network (segmentation-based approach) provides closer estimates to radiologist's annotations than the 3D-FasterCNN network. For the HPU network we also evaluated how relatively far the estimated nodule growth distribution was respect to the radiologist ground truth. To do this, we computed the Mahalanobis distance (and its associated probability) between the radiologist nodule growth annotation and the centre of the estimated diameter growth distribution. Additionally, we computed the probability of being the radiologist annotations, 1 or 2 standard deviations away from the centre of the estimated nodule growth distribution. Table 7. HPU nodule growth uncertainty performance. summarizes these results, in which the values represent the mean and 2 standard errors associated with a 95% of confidence, obtained from 1000 bootstraps with replacement of the test set.

Table 7. HPU nodule growth uncertainty performance.

	Test
Mahalanobis distance (RX_growth, distribution)	0.383 +/- 0.205
P(RX_growth) to the mean (of the distribution)	0.646 +/- 0.08
P(RX_growth) <= 1 std to the mean (of the distribution)	0.874 +/- 0.117
P(RX_growth) <= 2 std to the mean (of the distribution)	1.0 +/- 0.0

Figure 6. Comparative of radiologist growth measurements with results from nodule detector and nodule segmentation.



Nodule malignancy classification

Three different methods (3DCNN-MAL, TS-3DCNN and TS-3DCNN-MAL) were evaluated on the resulting 32 matching lung nodules (65% of them cancerous) from the re-identification step. For the evaluation of this methods, we directly used the 3DCNN-MAL classifier on the evaluation cases, while for the other two classifiers, to avoid data leakage for this evaluation, we retrained them before being evaluated, using the training partition of the VHLung with a 10-fold cross-validation. Particularly for the 3DCNN-MAL classifier, as it outputs 3 probabilities, we assumed cancer prediction when this classifier reported as maximum probability either the category suspicious or malignancy.

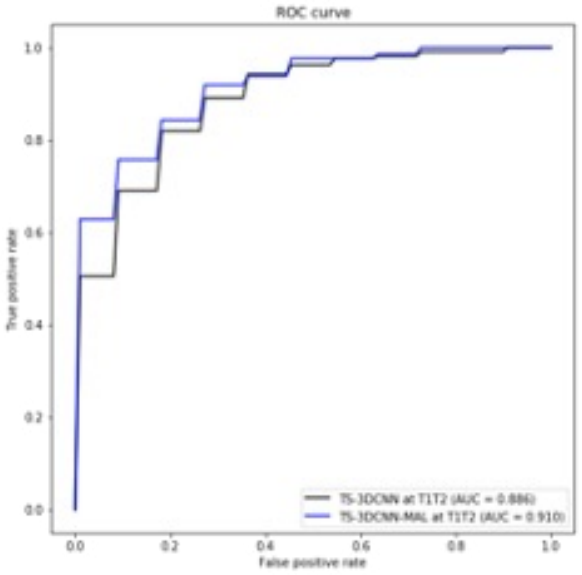
Table 8. Performance of the different nodule malignancy classifiers of component of the pipeline. shows the resulting classification performances for these models on the 32 matching nodules. We computed, precision (PREC), recall (REC) and specificity (SPEC), as well as balanced accuracy (BA). Due to the data was unbalanced towards the cancer case, the BA was used as the reference metric. Reported values in Table-4 are the mean and 2 standard errors (associated with a 95% of confidence).

Table 8. Performance of the different nodule malignancy classifiers of component of the pipeline.

	Test			
	BA	PREC	REC	SPEC
3DCNN-MAL	0.776+/-0.153	0.808+/-0.152	1.0+/-0.0	0.552+/-0.307
TS-3DCNN	0.810+/-0.203	0.874 +/- 0.117	0.874 +/- 0.117	0.874 +/- 0.117
TS-3DCNN-MAL	0.825+/-0.201	0.910+/-0.179	0.821+/-0.33	0.830+/-0.348

The 3DCNN-MAL classifier obtained a balanced accuracy score of 0.77, while the TS-3DCNN achieved a 0.81. However, the TS-3DCNN-MAL, which integrated the outcomes of the 3DCNN-MAL model, improved the balanced accuracy score of the TS-3DCNN model, a 1.5%. For further comparison of these models, we show Figure 7. ROC Curves of the nodule malignancy classification models.with the ROC-curves of the two best models.

Figure 7. ROC Curves of the nodule malignancy classification models.



4.3. CVC Up4Health-Characterization of Lung Pathologies

For all use cases, two different experiments are carried out:

1. Model Optimization. A training and selection of the best hyperparameters optimized using Optuna and a double k-fold stratification.
2. Model Verification. Validation of the best models on the independent set of test patients to assess the reproducibility of results and compare to state-of-art methods.

The metrics for assessing AI approaches are sensitivity, specificity, precision, positive predictive value (PPV), negative predictive value (NPV), diagnostic accuracy, F1-score, receiver operating characteristic (ROC) curves and the area under the curve (AUC). Multivariate analysis tests (Bonferroni-adjusted) are used to detect significant differences in quality scores across methods a p-value of 0.05 was considered significant.

4.3.1 Detection of Pulmonary Embolism in COVID19 pneumonia

In order to assess the impact of the size of the window used to compute radiomic features, we trained two different model with features extracted using windows of size, sze x sze = 3x3 and sze x sze = 5x5, to compute pyradiomics texture feature in a neighbourhood of each sample. To analyse the effect of the normalization of SPECT scans, for each window size a different model was trained with data normalized using the maximum and percentile criteria. A total number of 4 models were trained using the configurations reported in Table 9.

Table 9. Configuration of the different radiomic models

SPECT Normalization		
Window size	Max	Percentile
3 x 3	ModelMx3: SPECT normalization based on maximum values with 3x3 windows	ModelPrct3: SPECT normalization based on upper percentile values with 3x3 windows
5 x 5	ModelMx5: SPECT normalization based on maximum values with 5x5 windows	ModelPrct5: SPECT normalization based on upper percentile values with 5x5 windows
0.825+/-0.201	0.910+/-0.179	0.821+/-0.33

Table 10 reports statistical summary of the AUC score for the training set used for selection of models. For each radiomic model (rows) and diagnosis (columns), we report the AUC, its 95% confidence intervals (CI) and p-values for the comparison of AUC. The blank cells with * correspond to the best AUC that has been used to make the comparison. For, both, PE and pneumonia diagnosis AUC of ModelPrct3 was above 0.91, which confers percentile-based models with high diagnostic value.

Table 10. AUC Statistics for the Model Optimization

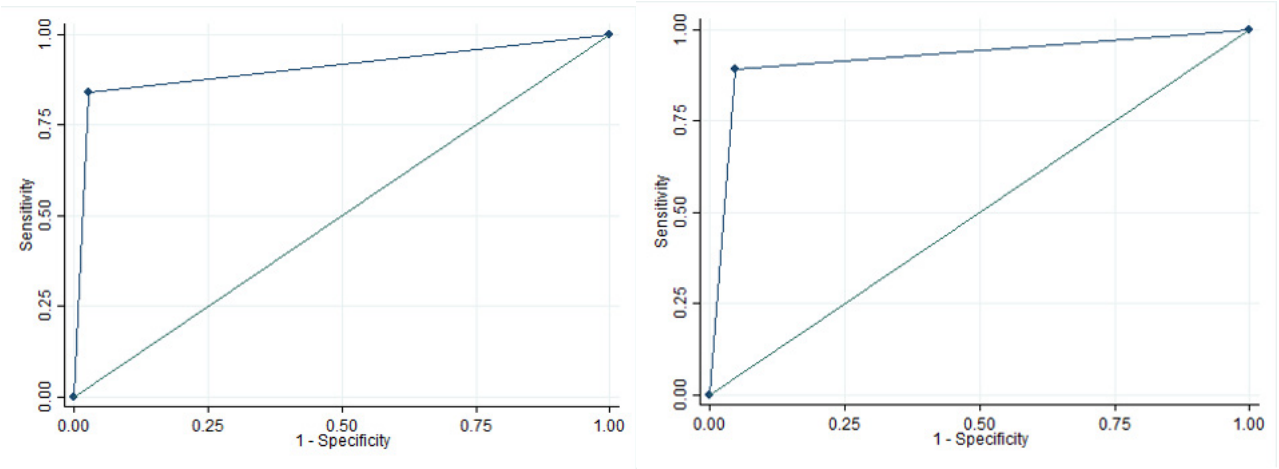
	SPECT Normalization			SPECT Normalization		
Radiomic model	AUC	95% CI	P-value	AUC	95% CI	P-value
ModelPrct3	0.919	0.909-0.929	*	0.922	0.914-0.931	*
ModelPrct5	0.902	0.892-0.913	0.028	0.905	0.896-0.914	<0.001
ModelMx3	0.835	0.822-0.848	<0.001	0.825	0.814-0.837	<0.001
ModelMx5	0.823	0.810-0.836	<0.001	0.818	0.806-0.830	<0.001

The AUC (Figure 8 Pulmonary Embolism and Pneumonia ROC area in Test set) for ModelPrct3 in test group is 0.922 (95% CI 0.090-0.935) for PE and 0.906 (95% CI 0.896-0.916) for Pneumonia, which proves that the model predictions are reproducible. Table 11 reports the sensitivity, specificity, positive predictive value, and negative predictive value for ModelPrct3, and we report average values and 95% confidence intervals (CI) of each score. Values for pneumonia diagnosis are almost the same as for the training set. Values for PE diagnosis are also comparable as for the training set, except for sensitivity, which is lower (75%). Since PPV is preserved as high as in training (88.9%), we consider this validates our model for PE diagnosis.

Table 11. Sensitivity, specificity, PPV and NPV statistics for the test set of ModelPrct3

	Pulmonary Embolism	Pneumonia
Sensitivity	75.1% (95% CI: 71.8-78.2)	93.3% (95% CI: 91.8-94.6)
Specificity	98.2% (95% CI: 97.7-98.6)	93.0% (95% CI: 92.1-93.9)
Positive predictive value	88.9% (95% CI: 86.2-91.2)	83.9% (95% CI: 81.8-85.7)
Negative predictive value	95.4% (95% CI: 94.7-96.0)	97.3% (95% CI: 92.3-93.8)

Figure 8 Pulmonary Embolism and Pneumonia ROC area in Test set



4.3.2 Diagnosis of Nodule Malignancy for Lung Cancer Screening

For this use case, the model optimization step included the optimization of the feature selection strategy (reproducibility, correlation or none) and CNN architecture.

For model optimization, 51 (85%) patients of the dataset described in the deliverable 1 *Acces to data* were randomly selected for the optimization of models. This training set had 8 benign and 43 malignant nodules. Table 12. Diagnosis scores of best models. reports sensitivity, specificity and F-score for malignancy classification for the different architectures considered with top performance highlighted in boldface. For all architectures, the proposed embedding (corresponding to Models 3, 6, 9 and 12) is the one that achieves better metrics with 100% of diagnostic sensitivity and specificity. Among them, the one with highest F-score is Model3, which is the one with the simplest architecture.

Table 12. Diagnosis scores of best models.

Model	Sensitivity	Specificity	F-score
Model1	100	100	0.856
Model2	93.2	75	0.683
Model3	100	100	0.903
Model4	100	87.5	0.846
Model5	97.67	37.5	0.595
Model6	100	100	0.839
Model7	100	100	0.804
Model8	100	37.5	0.619
Model9	100	100	0.834
Model10	100	87.5	0.840
Model11	100	37.5	0.617
Model12	100	100	0.831

In order to statistically evaluate the reproducibility our system, we have conformed an independent set of test patients from our database and the LIDC-IDRI public database. From our database we used 1 benign and 8 malignant nodules. Regarding the LIDC-IDRI database, since it was not collected to evaluate malignancy, scans are, in general, of a too low quality to assess malignancy. Following [CVC18, CVC19] we selected cases fulfilling the minimum acquisition requirements to allow radiological assessment of malignancy, which are slice thickness <= 2.5, resolution <= 0.71, except in case thickness is <= 1.5, that resolution can be <= 0.86 and only taking in consideration those nodules that have been diagnosed through a biopsy as benign or malign. After this filtering, 18 cases with diagnosis (5 benign and 13 malign) were selected. In this way, the independent set of test patients is formed by a total amount of 27 nodules with 6 benign and 21 malign.

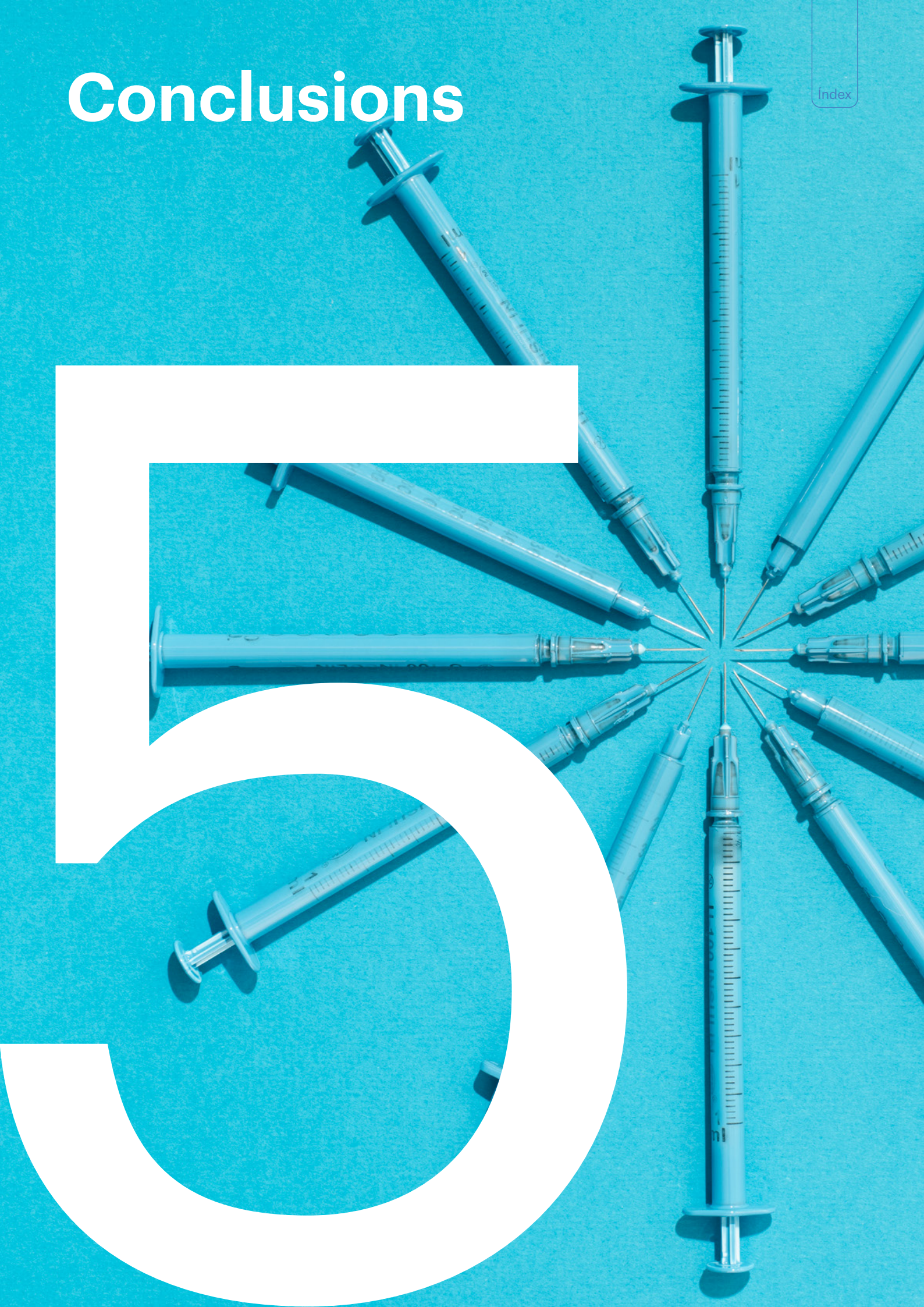
We have compared our Model3 with state-of-art methods which include three type of approaches: radiomics [CVC21], machine learning [CVC22] and deep CNN [CVC4, CVC5, CVC6, CVC7, CVC20]. Table 13. Results of our method compared to state-of-art. shows the metrics for state-of-art methods grouped according to type of approach and our method with best performance in boldface. It reports the metrics obtained by Model3 in our test set together with the results obtained by the selected state of the art in their datasets and reported in their works. We also report the number of parameters of each method as indicator of its complexity and computational and data cost for training. Our method outperforms in Accuracy, Sensitivity and F1 Score. In computer-aided diagnose, sensitivity is significant because correctly finding out patients with malignant nodules is crucial. Besides, the highest F1 Score implies that our method achieves the best trade-off between precision and recall. Our method has a splendid compromise between the performance of the system and the number of trainable parameters. A remarkable point compared to Deep CNN approaches, is that, our method needs strongly less samples to train the model, which is a must in medical imaging.

Table 13. Results of our method compared to state-of-art.

Approaches	Acc.	Sens.	Specif.	F1 Score	AUC	Param (M)
Radiomics	-	90.80	85.50	-	0.939	<0.29
Peikert et al. [CVC21]						
Machine Learning	96.09	96.84	95.34	-	0.979	<0.29
Zhang et al. [CVC8]						
Deep CNN	87.14	77.00	93.00	-	0.930	-
Multicrop [CVC20]	87.30	88.50	86.00	87.23	0.937	-
Nodule-level 2D [CVC4]	87.40	89.40	85.20	87.25	0.947	-
Vanilla 3D [CVC4]	90.44	81.42	-	-	-	141.57
DeepLung [CVC5]	90.24	92.04	88.94	90.45	0.933	678.69
AE-DPN [CVC6]	90.77	85.37	95.04	89.04	-	16.84
NASLung [CVC7]						
Hybrid	96.30	100	83.33	97.67	0.940	0.29
model3 (Our)						

Conclusions

5



In this deliverable document we presented the evolution, implementation and validation of the AI models developed by three use cases of elected for the PAI – CIDAI. Using this use cases as good practices in the application to the medical field, we propose a final guide focused in the variety of options to choose when setting up a new IA medical project. These conclusions are organised into the same different identified steps summarized in this document:

1. Infrastructure set-up: different strategies for data processing can be used. In one side, standalone applications require high resources in both, licencing costs and computation. On the other side, following the presented use cases, distributed computing in the cloud could be an alternative. These strategies could combine offline computations in the cloud as well as, online computations for real time processing. Cloud architectures could allow both, clinicians to access and use new IA methods and, researchers develop better robustness methods by accessing to new cases.
2. Deployment, execution and fine tuning of AI models: the development of IA methods have 5 different main steps: 1) Data Gathering; 2) Data Pre-processing; 3) Feature extraction; 4) Classification; and 5) Hyperparameter Optimization. The presented use cases apply these steps into a deep approach. The use of CNNs have a benefit in 2 ways. First, steps 3 and 4 are train by the network itself in one shot and, second, multimodal data is processed and mixed by the CNN. Final step, the hyperparameter optimization, will produce the best deep architecture configuration.
3. Model validation: once the optimal model is developed, final step is the validation of the methods. In medical image is commune to use different metrics to evaluate the system: Sensitivity, Specificity, Accuracy, F1-score and the ROC and AUC curves. In order to account for robustness, these metrics are applied to a independent test set for the final verification.

References

[CVC1] Ramírez, E., Sanchez, C., & Gil, D. (2021). BronchoX-Cloud: A Diagnostic as a Service Cloud Platform for Planning and Navigation Bronchoscopy. In 2021 23rd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC) (pp. 227-233). IEEE.

[CVC2] Degenhardt, F.; Seifert, S.; Szymczak, S. Evaluation of variable selection methods for random forests and omics data sets. Briefings in bioinformatics 2019, 20, 492–503.

[CVC3] Cohen, J.P.; Luck, M.; Honari, S. Distribution matching losses can hallucinate features in medical image translation. International conference on medical image computing and computer-assisted intervention. Springer, 2018, pp. 529–536.

[CVC4] Yan, X.; Pang, J.; Qi, H.; Zhu, Y.; Bai, C.; Geng, X.; Liu, M.; Terzopoulos, D.; Ding, X. Classification of lung nodule malignancy risk on computed tomography images using convolutional neural network: A comparison between 2d and 3d strategies. Asian Conference on Computer Vision. Springer, 2016, pp. 91–101.

[CVC5] Zhu, W.; Liu, C.; Fan, W.; Xie, X. Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018, pp. 673–681.

[CVC6] Jiang, H.; Gao, F.; Xu, X.; Huang, F.; Zhu, S. Attentive and ensemble 3D dual path networks for pulmonary nodules classification. Neurocomputing 2020, 398, 422–430.

[CVC7] Jiang, H.; Shen, F.; Gao, F.; Han, W. Learning efficient, explainable and discriminative representations for pulmonary nodules classification. Pattern Recognition 2021, 113, 107825.

[CVC8] Zhang, F.; Song, Y.; Cai, W.; Lee, M.Z.; Zhou, Y.; Huang, H.; Shan, S.; Fulham, M.J.; Feng, D.D. Lung nodule classification with multilevel patch-based context analysis. IEEE Transactions on Biomedical Engineering 2013, 61, 1155–1166.

[CVC9] Maicas, G.; Bradley, A.P.; Nascimento, J.C.; Reid, I.; Carneiro, G. Training medical image analysis systems like radiologists. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2018, pp. 546–554.

[CVC10] Gil, D., Baeza, S., Sanchez, C., Torres, G., García-Olivé, I., Moragas, G., ... & Rosell, A. (2021). Intelligent Radiomic Analysis of Q-SPECT/CT images to optimize pulmonary embolism diagnosis in COVID-19 patients. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 446-453).

[CVC11] Torres, G., Baeza, S., Sanchez, C., Guasch, I., Rosell, A., & Gil, D. (2022). An Intelligent Radiomic Approach for Lung Cancer Screening. Applied Sciences, 12(3), 1568.

[CVC12] Hernández-Sabaté, A., Yauri, J., Folch, P., Piera, M. À., & Gil, D. (2022). Recognition of the Mental Workloads of Pilots in the Cockpit Using EEG Signals. Applied Sciences, 12(5), 2298.

[CVC13] Baeza, S., Domingo, R., Salcedo-Pujantell, M., Deportós, J., Moragas, G., Garcia-Olivé, I., ... & Rosell, A. (2021). Artificial intelligence to optimize pulmonary embolism diagnosis during covid-19 pandemic by perfusion SPECT/CT, a pilot study.

[CVC14] Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2019).

[CVC15] Varma, S., Simon, R.: Bias in error estimation when using cross-validation for model selection. BMC bioinformatics 7(1), 1–8 (2006).

[CVC16] Beig, N.; Khorrami, M.; Alilou, M.; Prasanna, P.; Braman, N.; Orooji, M.; Rakshit, S.; Bera, K.; Rajiah, P.; Ginsberg, J.; others. Perinodular and intranodular radiomic features on lung CT images distinguish adenocarcinomas from granulomas. Radiology 2019, 290, 783–792.

[CVC17] Gil, D.; Sanchez, C.; Borras, A.; Diez-Ferrer, M.; Rosell, A. Segmentation of distal airways using structural analysis. Plos one 2019, 14, e0226006.

[CVC18] Kim, Y.J.; Lee, H.J.; Kim, K.G.; Lee, S.H. The effect of CT scan parameters on the measurement of CT radiomic features: a lung nodule phantom study. Computational and mathematical methods in medicine 2019, 2019.

[CVC19] Xu, Y.; Lu, L.; Sun, S.H.; Lian, W.; Yang, H.; Schwartz, L.H.; Yang, Z.h.; Zhao, B.; others. Effect of CT image acquisition parameters on diagnostic performance of radiomics in predicting malignancy

of pulmonary nodules of different sizes. European Radiology 2021, pp. 1–11.

[CVC20] Shen, W.; Zhou, M.; Yang, F.; Yu, D.; Dong, D.; Yang, C.; Zang, Y.; Tian, J. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. Pattern Recognition 2017, 61, 663–673.

[CVC21] Peikert, T., Duan, F., Rajagopalan, S., Karwoski, R. A., Clay, R., Robb, R. A., ... & Maldonado, F. (2018). Novel high-resolution computed tomography-based radiomic classifier for screen-identified pulmonary nodules in the National Lung Screening Trial. PLoS One, 13(5), e0196910.

[EUT1] Pelc, Norbert J. "Recent and future directions in CT imaging." Annals of biomedical engineering 42.2 (2014): 260-268.

[EUT 2] National Lung Screening Trial Research Team. "Reduced lung-cancer mortality with low-dose computed tomographic screening." New England Journal of Medicine 365.5 (2011): 395-409.

[EUT3] Bashir, Usman, et al. "Imaging heterogeneity in lung cancer: techniques, applications, and challenges." American Journal of Roentgenology 207.3 (2016): 534-543.

[EUT4] Aziz, Z. A., S. P. Padley, and D. M. Hansell. "CT techniques for imaging the lung: recommendations for multislice and single slice computed tomography." European journal of radiology 52.2 (2004): 119-136.

[EUT5] Han, Paul KJ, William MP Klein, and Neeraj K. Arora. "Varieties of uncertainty in health care: a conceptual taxonomy." Medical Decision Making 31.6 (2011): 828-838.

[EUT6] Kaya, Aydın, and Ahmet Burak Can. "A weighted rule based method for predicting malignancy of pulmonary nodules by nodule characteristics." Journal of biomedical informatics 56 (2015): 69-79.

[EUT7] Suzuki, Kenji. "Machine learning in computer-aided diagnosis of the thorax and colon in CT: a survey." IEICE transactions on information and systems 96.4 (2013): 772-783.

[EUT8] Coroller, Thibaud P., et al. "CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma." Radiotherapy and Oncology 114.3 (2015): 345-350.

[EUT9] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 436-444.

[EUT10] Esteva, Andre, et al. "Dermatologist-level classification of skin cancer with deep neural networks." nature 542.7639 (2017): 115-118.

[EUT11] Gulshan, Varun, et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs." Jama 316.22 (2016): 2402-2410.

[EUT12] Bejnordi, Babak Ehteshami, et al. "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer." Jama 318.22 (2017): 2199-2210.

[EUT13] Setio, Arnaud Arindra Adiyoso, et al. "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge." Medical image analysis 42 (2017): 1-13.

[EUT14] Messay, Temesguen, Russell C. Hardie, and Timothy R. Tuinstra. "Segmentation of pulmonary nodules in computed tomography using a regression neural network approach and its application to the lung image database consortium and image database resource initiative dataset." Medical image analysis 22.1 (2015): 48-62.

[EUT15] Ciompi, Francesco, et al. "Towards automatic pulmonary nodule management in lung cancer screening with deep learning." Scientific reports 7.1 (2017): 1-11.

[EUT16] Bonavita, Ilaria, et al. "Integration of convolutional neural networks for pulmonary nodule malignancy assessment in a lung cancer classification pipeline." Computer methods and programs in biomedicine 185 (2020): 105172.

[EUT17] Rafael-Palou, Xavier, et al. "Re-Identification and growth detection of pulmonary nodules without image registration using 3D siamese neural networks." Medical Image Analysis 67 (2021): 101823.

[EUT18] Rafael-Palou, Xavier, et al. "Pulmonary Nodule Malignancy Classification Using its Temporal Evolution with Two-Stream 3D Convolutional Neural Networks." Medical Imaging with Deep Learning. 2020.

[EUT19] Ding, Jia, et al. "Accurate pulmonary nodule detection in computed tomography images using

deep convolutional neural networks." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2017.

[EUT20] Xie, Hongtao, et al. "Automated pulmonary nodule detection in CT images using deep convolutional neural networks." Pattern Recognition 85 (2019): 109-119.

[EUT21] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." arXiv preprint arXiv:1506.01497 (2015).

[EUT22] Liao, Fangzhou, et al. "Evaluate the malignancy of pulmonary nodules using the 3-d deep leaky noisy-or network." IEEE transactions on neural networks and learning systems 30.11 (2019): 3484-3495.

[EUT23] Zhu, Wentao, et al. "Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification." 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018.

[EUT24] Kuan, Kingsley, et al. "Deep learning for lung cancer detection: tackling the kaggle data science bowl 2017 challenge." arXiv preprint arXiv:1705.09435 (2017).

[EUT25] Schwartz, Mordecai. "A biomathematical approach to clinical tumor growth." Cancer 14.6 (1961): 1272-1294.

[EUT26] Han, Daiwei, et al. "Influence of lung nodule margin on volume-and diameter-based reader variability in CT lung cancer screening." The British journal of radiology 91.1090 (2018): 20170405.

[EUT27] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.

[EUT28] Çiçek, Özgün, et al. "3D U-Net: learning dense volumetric segmentation from sparse annotation." International conference on medical image computing and computer-assisted intervention. Springer, Cham, 2016.

[EUT29] Milletari, Fausto, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation." 2016 fourth international conference on 3D vision (3DV). IEEE, 2016.

[EUT30] Isensee, Fabian, et al. "nnu-net: Self-adapting framework for u-net-based medical image segmentation." arXiv preprint arXiv:1809.10486 (2018).

[EUT31] Larici, Anna Rita, et al. "Lung nodules: size still matters." European Respiratory Review 26.146 (2017).

[EUT32] Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles." arXiv preprint arXiv:1612.01474 (2016).

[EUT34] Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." The journal of machine learning research 15.1 (2014): 1929-1958.

[EUT33] Kendall, Alex, Vijay Badrinarayanan, and Roberto Cipolla. "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding." arXiv preprint arXiv:1511.02680 (2015).

[EUT34] Sohn, Kihyuk, Honglak Lee, and Xinchen Yan. "Learning structured output representation using deep conditional generative models." Advances in neural information processing systems 28 (2015): 3483-3491.

[EUT35] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[EUT36] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

[EUT37] Siegelman, Stanley S., et al. "Solitary pulmonary nodules: CT assessment." Radiology 160.2 (1986): 307-312.

[EUT38] MacMahon, Heber, et al. "Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017." Radiology 284.1 (2017): 228-243.

[EUT39] Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[EUT40] Kohl, Simon AA, et al. "A hierarchical probabilistic u-net for modeling multi-scale ambiguities." arXiv preprint arXiv:1905.13077 (2019).

[EUT41] Sun, Wenqing, Bin Zheng, and Wei Qian. "Computer aided lung cancer diagnosis with deep learning algorithms." Medical imaging 2016: computer-aided diagnosis. Vol. 9785. International Society for Optics and Photonics, 2016.

[EUT42] Causey, Jason L., et al. "Highly accurate model for prediction of lung nodule malignancy with CT scans." Scientific reports 8.1 (2018): 1-12.

[EUT43] Dey, Raunak, Zhongjie Lu, and Yi Hong. "Diagnostic classification of lung nodules using 3D neural networks." 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). IEEE, 2018.

[EUT44] Ardila, Diego, et al. "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography." Nature medicine 25.6 (2019): 954-961.

[EUT45] Veasey, Benjamin P., et al. "Lung Nodule Malignancy Prediction From Longitudinal CT Scans With Siamese Convolutional Attention Networks." IEEE Open Journal of Engineering in Medicine and Biology 1 (2020): 257-264.

[EUT46] van Riel, Sarah J., et al. "Observer variability for classification of pulmonary nodules on low-dose CT images and its effect on nodule management." Radiology 277.3 (2015): 863-871.

[EUT47] Shrivastava, Abhinav, Abhinav Gupta, and Ross Girshick. "Training region-based object detectors with online hard example mining." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[EUT48] Koch, Gregory, Richard Zemel, and Ruslan Salakhutdinov. "Siamese neural networks for one-shot image recognition." ICML deep learning workshop. Vol. 2. 2015.

[EUT49] Armato III, Samuel G., et al. "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans." Medical physics 38.2 (2011): 915-931.

[EUT50] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017.

[EUT51] Zhang, Wei, et al. "Parallel distributed processing model with local space-invariant interconnections and its optical architecture." Applied optics 29.32 (1990): 4790-4797.

[EUT52] Huang, Peng, et al. "Prediction of lung cancer risk at follow-up screening with low-dose CT: a training and validation study of a deep learning method." The Lancet Digital Health 1.7 (2019): e353-e362.

[EUT53] Tammemagi, Martin C., et al. "Participant selection for lung cancer screening by risk modelling (the Pan-Canadian Early Detection of Lung Cancer [PanCan] study): a single-arm, prospective study." The lancet oncology 18.11 (2017): 1523-1531.

[EUT54] Bello, Irwan, et al. "Attention augmented convolutional networks." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[EUT55] Parmar, Niki, et al. "Image transformer." International Conference on Machine Learning. PMLR, 2018.

[EUT56] Revel, Marie-Pierre, et al. "Are two-dimensional CT measurements of small noncalcified pulmonary nodules reliable?." Radiology 231.2 (2004): 453-458.

[EUT57] Huang, Wenkai, Yihao Xue, and Yu Wu. "A CAD system for pulmonary nodule prediction based on deep three-dimensional convolutional neural networks and ensemble learning." Plos one 14.7 (2019): e0219369.

[EUT58] Song, Guoli, et al. "A review on medical image registration as an optimization problem." Current Medical Imaging 13.3 (2017): 274-283.

[EUT59] Goodfellow, Ian J., et al. "Generative adversarial networks." arXiv preprint arXiv:1406.2661 (2014).

[EUT60] Goodfellow, Ian. "Nips 2016 tutorial: Generative adversarial networks." arXiv preprint arXiv:1701.00160 (2016).

CIDAI Centre of Innovation for Data tech and Artificial Intelligence

C/Bilbao, 72 Edif. A.
08005 - Barcelona
Tel. +34 93 7419 100
info@cidai-catalonia-ai.eu
www.cidai.eu