GDAŃSK UNIVERSITY OF TECHNOLOGY

TASK Quarterly 30 (2) 2026

Systematic Evaluation of 2D and 3D Deep Learning Detectors for Pulmonary Nodule Detection in Chest CT

Milena Sobotka, Antoni Górecki, Tomasz Neumann, Konrad Mrozowski, Dmytro Tkachenko, Natalia Kowalczyk, Magdalena Mazur Milecka*, Tomasz Kocejko, Jacek Rumiński

Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, Narutowicza 11/12, 80-233 Gdańsk, Poland

*magmilec@pg.edu.pl

https://doi.org/10.34808/tq2026/30.2/a

Abstract

Accurate detection of pulmonary nodules in chest CT is essential for early lung cancer diagnosis. This study presents a systematic evaluation of two-dimensional (2D) and three-dimensional (3D) deep learning detectors for pulmonary nodule detection using the LUNA16 benchmark dataset. While 3D deep learning approaches can exploit full volumetric information, 2D detectors remain attractive due to their lower computational complexity and simpler training requirements. Particular attention was devoted to the effect of annotation preprocessing in the 2D setting.

Two 2D detectors (RetinaNet and YOLO) and two 3D detectors (3D RetinaNet and a hybrid CNN-Transformer architecture) were evaluated. For the 2D experiments, multiple dataset variants were prepared by filtering annotations according to minimum bounding-box area and by balancing positive and negative slices. Detection performance was evaluated using F1-score, precision, recall, mAP, mAR, and FROC-based metrics across multiple IoU thresholds. The results showed that preprocessing strategies substantially affected 2D detector performance. Filtering very small peripheral annotations and balancing the datasets improved localization quality and detector stability for both RetinaNet and YOLO. Among the 2D configurations, the best results were obtained for the balanced filtered subsets.

The 3D RetinaNet achieved the strongest overall performance, providing higher detection sensitivity, more stable localization, and better generalization between validation and independent test data. The CNN-Transformer also benefited from volumetric input and retained relatively high recall at lower IoU thresholds, although its performance remained lower than that of the 3D RetinaNet.

The findings demonstrate the importance of volumetric context for robust pulmonary nodule detection and indicate that appropriate annotation preprocessing can substantially improve the effectiveness of slice-based 2D detection frameworks.

Keywords:

pulmonary nodule detection, chest CT, 2D vs 3D detection

1. Introduction

Pulmonary nodule detection in chest CT is an important task in computer-aided diagnosis, as early identification of suspicious nodules may support lung cancer screening and clinical decision-making. Deep learning methods have become widely used for this purpose, but the choice between two-dimensional slice-based detectors and three-dimensional volumetric detectors remains an important practical issue. While 2D methods are easier to train and can use established object detection architectures, 3D methods preserve volumetric context at the cost of greater computational complexity.

In this work, we address this gap by conducting a comprehensive and controlled comparison of representative two-dimensional and three-dimensional object detection models for pulmonary nodule detection using the LUNA16 dataset. We construct consistent two-dimensional and three-dimensional representations of the data, including filtered variants designed to account for the detectability of small lesions, and evaluate multiple detector families, including RetinaNet and YOLO in 2D, as well as a MONAI-based 3D RetinaNet and a hybrid 3D CNN–Transformer architecture. By using a common benchmark, a shared reference standard, patient-level data splitting, and consistent detection metrics within each representation, this study provides a controlled analysis of the practical differences between 2D slice-based and 3D volumetric approaches. Since 2D and 3D detectors require different input representations and annotation formats, the comparison should be interpreted as a representation-level performance and efficiency analysis rather than a strictly identical-sample training comparison.

2. Related Work

The detection of pulmonary nodules in CT images is one of the challenging applications of computer-aided diagnostic imaging. CT data are volumetric, meaning that a single scan consists of a sequence of image slices that together represent a three-dimensional volume of the examined region. Information that may aid diagnosis is not always visible in a single slice, but can be spread across many slices. Small lesions, such as pulmonary nodules, can occupy only a few of the hundreds or dozens of slices, have low contrast, or resemble other anatomical structures within the examined area. This task is a complex machine learning problem, as lesions are often small relative to the section's overall volume, leading to an imbalance in the ratio of the region of interest to the background.

Many anatomical structures in CT scans may resemble pulmonary nodules, which increases the number of false positives. In addition, preparing training data often requires expertise from medical experts, and accurately annotating spatial locations is time-consuming and costly. These limitations have led to the development of methods that use bounding boxes, diameters, and center points in place of complete and accurate segmentation masks.

2.1. 2D detection models

One approach is to apply conventional object detection methods by treating each CT slice as a separate image. One of the major advantages of this approach is the ability to leverage well-known, thoroughly tested computer vision model architectures, such as Faster R-CNN, RetinaNet, Mask R-CNN, or the YOLO model family.

In the context of 2D object detection, it is worth noting the Faster R-CNN architecture, proposed by Ren, He, Girshick and Sun [1]. The system consists of two main modules: a convolutional network that generates region proposals, and the Fast R-CNN architecture, which uses these proposals for object classification and bounding box regression [1]. A key innovation of this approach was the integration of the region proposal generation process with the detection network, which enabled the sharing of feature maps and significantly accelerated the operation of the entire model [1].

This approach has also been applied to the task of detecting lung nodules in computed tomography (CT) images. In this domain, Ding, Li, Hu, and Wang proposed extending Faster R-CNN with a deconvolution layer, designed to better capture the fine features characteristic of lung nodules [2]. The authors noted that the original Faster R-CNN network, which utilized five groups of VGG-16 convolutional layers for feature extraction, was not sufficiently effective at detecting small regions of interest corresponding to nodules [2]. They also noted that using the full three-dimensional CT scan volume as input to the DCNN entails high computational costs. For this reason, axial slices, supplemented with information from neighbouring layers, were used as input instead of the full 3D volume [2]. The proposed solution achieved a sensitivity of 94.6% with 15 candidates per scan in the candidate detection stage, and the entire system achieved an average FROC score of 0.891 on the LUNA16 benchmark [2].

Faster R-CNN and systems built on this architecture utilized a two-stage approach. However, among single-stage detectors, the YOLO family of models is of particular significance. The main idea of the authors Redmon, Divvala, Girshick, and Farhadi was to transform object detection into a single regression task, in which the grid cells superimposed on the image are directly assigned the coordinates of bounding boxes and the probabilities of class

membership [3]. A single convolutional network simultaneously predicts multiple bounding boxes and the class membership probabilities for those boxes. Each bounding box consists of 5 values: x, y, w, h and a confidence score. The coordinates (x, y) indicate the center of the bounding box relative to the grid cell boundaries. The width and height are predicted relative to the entire image [3].

Models from the YOLO family were subsequently adapted for lung nodule detection in CT scans. An example of such an adaptation is the YOLO-MSRF method, proposed by Wu, Zhang, Sun, Wang, and Zhang, based on the YOLOv7 architecture [4]. The authors noted that the process of pre-generating candidate frames is very time-consuming and slows the network's detection performance; furthermore, they observed that two-stage detector-based approaches require prior knowledge, which can lead to inaccurate localization of candidate frames. The authors mentioned that detection systems using a two-stage approach achieve acceptable results, yet certain problems remain, such as: low detection accuracy or the inability to fully utilize information about the features of the object's surroundings [4]. Therefore, building upon the then-current YOLOv7, the authors extended the model with an additional layer for detecting small objects, a multi-scale receptive field module, and efficient omni-dimensional convolution. This was done to improve the detection of small and irregular pulmonary nodules [4]. The method was evaluated on the public LUNA16 dataset and achieved mAP, precision, and recall values of 95.26%, 95.41%, and 94.02%, respectively [4].

2.2. 3D detection models

A recent and increasingly explored research direction involves hybrid architectures that combine 3D convolutional neural networks with transformer-based modules. These models aim to integrate the strong local feature extraction capabilities of convolutional operators with the ability of attention mechanisms to capture long-range dependencies. In the context of pulmonary nodule detection, this approach is particularly relevant, as very small lesions may require both fine-grained local texture and shape information, as well as broader anatomical context from surrounding structures. An example of such an approach is TiCNet, which embeds transformer modules within a 3D CNN framework for lung nodule detection [5]. Transformer-based models such as Vision Transformers have demonstrated strong capabilities in modeling global context in medical imaging [6].

From the perspective of comparing two-dimensional and three-dimensional detection approaches, the current literature does not provide a definitive conclusion that three-dimensional models consistently outperform their two-dimensional counterparts. While several studies demonstrate that 3D models can better exploit volumetric spatial context, others indicate that, in practical pipelines, the performance gains may be relatively modest and come at a significantly higher computational cost. For example, Setio et al. showed that multi-view and 3D context improves false positive reduction in lung nodule detection, but at the cost of increased model complexity [7]. Similarly, Dou et al. demonstrated that 3D CNNs outperform 2D counterparts for false positive reduction, although the improvement is limited relative to computational demands [8].

These observations highlight the need for systematic and controlled comparisons of two-dimensional and three-dimensional detection paradigms under consistent experimental conditions. In particular, meaningful evaluation requires the use of a common dataset, a unified definition of target lesions, and standardized evaluation protocols. Such a setup enables a more reliable assessment of the trade-offs between detection performance, robustness, and computational efficiency.

3. Methods

This section describes the datasets, preprocessing procedures, annotation representations, and detection models used in the experiments. The methodology was designed to enable a consistent comparison between 2D slice-based detectors and 3D volumetric detectors.

3.1. Datasets

All experiments were conducted using the LUNA16 benchmark dataset. The dataset was prepared in separate 2D and 3D representations in order to support both slice-based and volumetric detection pipelines.

3.1.1 LUNA16

LUNA16 is a publicly available benchmark for pulmonary nodule detection derived from the LIDC-IDRI database [9]. The dataset contains 888 thoracic CT scans selected from the original 1018 LIDC-IDRI studies after excluding scans with excessive slice thickness and scans with inconsistent slice spacing or missing slices. The scans are distributed in MetaImage format and are organized into ten official subsets intended for ten-fold cross-validation. [10–12]

The underlying LIDC-IDRI annotations were produced by four experienced thoracic radiologists in a two-stage reading process. During the blinded phase, each reader independently marked lesions as nodules > 3 mm, nodules < 3 mm, or non-nodules; in the

subsequent unblinded phase, the readers reviewed each other's findings without forced consensus. In LUNA16, only nodules with diameter of at least 3 mm that were accepted by at least three of the four radiologists are treated as positive lesions in the reference standard. This results in 1186 reference nodules over 888 scans. Findings marked by fewer than three readers, nodules smaller than 3 mm, and non-nodule abnormalities are treated as irrelevant findings and are excluded from the positive reference set. [10, 11]

In this study, LUNA16 was chosen as the sole benchmark in order to ensure a controlled comparison between 2D and 3D detectors under a common reference standard, identical patient-level splits, and clinically interpretable detection metrics. [10, 12]

3.1.2 LUNA16 2D version

To enable the use of 2D object detectors, the original LUNA16 CT volumes were converted into individual axial slices. This representation allowed the application of standard 2D detection models, such as YOLO and RetinaNet, which operate on single images rather than full volumetric data. Nodule annotations were obtained from the `annotations.csv` file provided with the LUNA16 dataset. The file contains the three-dimensional center coordinates of each lesion, denoted as X, Y, and Z, together with the nodule diameter expressed in millimeters [11].

The conversion from the volumetric annotation space to the two-dimensional image space was performed using the `origin` and `spacing` parameters stored in the MHD image headers. These parameters were used to transform the nodule coordinates into voxel indices and to determine the axial slices in which each nodule was present. For each selected slice, the visible part of the nodule was projected onto the XY plane, and a corresponding two-dimensional bounding box was generated. Its size depended on the distance between the given slice and the nodule center. The boxes were largest near the central slice of the lesion and became smaller on peripheral slices. The bounding-box dimensions were then converted into pixel units using the in-plane image spacing.

HU intensity windowing was applied to each slice before saving, using a window center of -600 HU and a window width of 1500 HU, in order to enhance contrast in the lung parenchyma region. The resulting 2D dataset, referred to as `LUNA_full`, comprised 127553 axial slices extracted from the LUNA16 CT volumes. Among them, 6089 slices contained at least one nodule annotation, with a total of 6342 bounding boxes, while 121464 slices were background images without any annotated lesion. The bounding-box areas ranged from 1.0 to 1650.2 px^2, with a mean of 194.37 px^2 and a median of 88.69 px^2. The data were split into training, validation, and test subsets in an approximate ratio of 70%/20%/10%. The split was performed at the CT-volume level, so that all 2D slices extracted from the same scan were assigned to the same subset, reducing the risk of data leakage between training, validation, and test data.

3.1.3 LUNA16 2D filtered variants

After converting LUNA16 into a two-dimensional representation, the distribution of the generated bounding-box areas was analyzed. Since the dataset contained a large number of very small bounding boxes, mainly corresponding to peripheral cross-sections of nodules, additional dataset variants were prepared by filtering annotations according to a minimum bounding-box area.

The filtering procedure was motivated by the characteristics of volumetric CT data and the limitations of slice-wise 2D representations. Pulmonary nodules are visible across multiple consecutive CT slices and often exhibit heterogeneous appearance throughout the volume. After conversion to independent 2D slices, peripheral sections of nodules frequently contain only a small visible fragment of the lesion, resulting in extremely small bounding boxes with limited anatomical context. Such annotations may introduce ambiguity and increase localization difficulty for 2D detectors, particularly for anchor-based architectures that are sensitive to very small objects. Therefore, filtering small annotations was intended to reduce annotation noise and improve the consistency of slice-level lesion representation.

The full dataset without annotation filtering was used as the baseline variant (`LUNA_full`). In addition, three filtered variants without negative-slice balancing were created: `LUNA_thr0.2`, `LUNA_thr1.0`, and `LUNA_thr2.4`. The `LUNA_thr0.2` variant used a minimum bounding-box area threshold of 38.87 px^2, which approximately corresponds to a 6×6 px bounding box. The `LUNA_thr1.0` variant used a threshold of 194.37 px^2, equal to the mean bounding-box area in the dataset and approximately corresponding to a 14×14 px box. The most restrictive unbalanced variant, `LUNA_thr2.4`, used a threshold of 466.49 px^2, corresponding approximately to a 22×22 px box. Depending on the selected threshold, the number of retained annotations and positive slices decreased substantially, allowing the influence of small peripheral annotations on model performance to be assessed.

For selected filtered configurations, the number of positive and negative slices was additionally balanced. This procedure consisted of adding the same number of slices without nodules as slices containing nodules, in order to reduce the strong

imbalance between positive and negative images that naturally occurs in CT data. The balanced variants were denoted as `LUNA_thr1.0_balanced` and `LUNA_thr2.4_balanced`. Overall, six dataset configurations were used in the 2D experiments: `LUNA_full`, `LUNA_thr0.2`, `LUNA_thr1.0`, `LUNA_thr1.0_balanced`, `LUNA_thr2.4`, and `LUNA_thr2.4_balanced`. Comparing these variants made it possible to evaluate how removing very small annotations and modifying the ratio between positive and negative slices affected the performance of 2D nodule detectors.

3.1.4 LUNA16 3D version

For the 3D detection experiments, 601 CT scans containing at least one annotated pulmonary nodule were used. The dataset was divided into training, validation, and test subsets. The training subset included 415 scans from subsets 1-7, the validation subset included 67 scans from subset 0, and the test subset included 119 scans from subsets 8 and 9. These subsets contained 851, 112, and 223 reference nodules, respectively. As part of preprocessing, all CT volumes were geometrically standardized. The source data stored in MHD/RAW format were converted to NIfTI, reoriented to the RAS coordinate system, and resampled to a fixed voxel spacing of $0.703125 \times 0.703125 \times 1.25$ mm. Nodule annotations were represented as three-dimensional bounding boxes.

3.2. Models

Pulmonary nodule detection in CT can be formulated as a localization problem, where the objective is to determine the presence of a lesion and estimate its spatial extent. This study considers four detection architectures: two 2D detectors (YOLO and RetinaNet) and two 3D detectors (3D RetinaNet and a hybrid 3D convolutional-transformer architecture), representing a range of contemporary approaches that differ in their ability to capture spatial context, from 2D slice-based representations to full volumetric modeling with global contextual integration.

3.2.1 2D detectors

For the 2D object detection experiments, RetinaNet and YOLOv12m were trained on the LUNA16 2D dataset and on its filtered variants described in section 3.1.3. All experiments were conducted on the same set of dataset versions in order to assess the effect of annotation filtering and balancing on detector performance.

3.2.2 3D detectors

RetinaNet for 3D

For volumetric detection, we used a 3D RetinaNet implemented with the MONAI detection framework. RetinaNet is a one-stage dense detector originally introduced for natural-image object detection and designed to address extreme foreground–background imbalance through focal loss [13].

In the MONAI implementation, the detector is built around a ResNet-50 backbone coupled with a feature pyramid network that extracts multi-scale feature maps from the input volume [14]. This design is particularly relevant for pulmonary nodule detection because nodules exhibit substantial scale variation, ranging from very small lesions spanning only a few slices to larger solid nodules. The detector operates directly on 3D inputs and predicts volumetric bounding boxes [15–17].

The detection head consists of two task-specific branches defined on top of the pyramid features: a classification head that outputs class logits for anchors at each spatial location, and a regression head that predicts box offsets.

During training, the MONAI RetinaNet detector takes input volumes together with target labels and volumetric ground-truth boxes, and optimizes classification and regression losses. During inference, predicted anchors are decoded into 3D boxes and post-processed using score thresholding, candidate filtering, and non-maximum suppression. In our study, this model serves as the canonical anchor-based 3D baseline, representing a strong convolutional detector that explicitly preserves inter-slice continuity and volumetric lesion geometry. [16, 17]

Hybrid 3D CNN–Transformer Architecture

The second 3D model is a hybrid CNN-Transformer detector that combines local volumetric feature extraction with global token-level contextualization. This design is motivated by the complementary strengths of convolutional operators for inductive locality and Vision Transformer attention for long-range dependency modeling in medical imaging contexts [18]. The architecture follows a two-stage design that combines convolutional feature extraction with transformer-based contextual modeling. The input to the network is a volumetric CT patch. The first stage consists of a three-dimensional convolutional encoder responsible for extracting local spatial features that progressively reduces spatial resolution while increasing the number of feature channels. This stage captures local anatomical structures such as edges, textures, and small volumetric patterns characteristic of pulmonary nodules. The use of 3D convolutions allows modeling of inter-slice continuity, which is not explicitly available in 2D approaches.

In the second stage, the feature tensor is subsequently transformed into a sequence of tokens suitable for transformer processing. This is achieved by: flattening the spatial dimensions, followed by a linear projection into an embedding space, adding positional embeddings as well as a class token. The token sequence is processed by a stack of transformer encoder blocks. The output corresponding to the class token is used as a global representation of the input volume. Two task-specific prediction heads operate on the global representation: (1) the classification branch predicts the probability of nodule presence, (2) the regression branch predicts the normalized coordinates of a 3D bounding box corresponding to the opposing corners of the lesion volume.
The model is trained using a multi-task loss:

$$L = L_{cls} + L_{box}. \tag{1}$$

The classification loss is defined as binary cross-entropy. For localization, we use mean-squared error between predicted and target box coordinates, applied only to positive samples through a label-based mask. A masking mechanism ensures that regression loss is not computed for negative patches that contain no target lesion.
To construct training data, scans are sampled into fixed-size 3D crops. Positive crops are sampled around annotated nodules, while negative crops are sampled randomly and rejected if they overlap any annotated lesion.

3.3. Evaluation metrics

To enable a fair comparison between 2D slice-based detectors and 3D volumetric detectors as transparent as possible, we distinguish between metrics used during model development (validation) and metrics used for final lesion-level assessment on the independent LUNA16 test split. Whenever possible, 3D models were evaluated on the same test scans and reference nodules.

3.3.1 2D metrics

For the 2D detectors (RetinaNet and YOLO), each predicted bounding box was compared with ground-truth boxes on the same axial slice. Detection quality was evaluated using the Intersection over Union (IoU) between predicted and reference two-dimensional boxes. A prediction was considered a true positive at threshold t if the corresponding IoU exceeded the selected threshold value. Metrics were evaluated for IoU thresholds in the range 0.10–0.50.

Average Precision (AP) and Average Recall (AR) were computed following the standard object-detection definition [19, 20]. We report AP and AR at IoU = 0.10 ($AP_{0.10}$ and $AR_{0.10}$), as well as mean values averaged across thresholds from 0.10 to 0.50 ($mAP_{0.10:0.50}$ and $mAR_{0.10:0.50}$), summarizing both localization quality and ranking behavior across different overlap criteria.

Classification quality at a fixed operating point was summarized using precision, recall, and the F1-score:

$$\text{Precision} = \frac{TP}{TP+FP}, \qquad \text{Recall} = \frac{TP}{TP+FN}, \tag{2}$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3}$$

Here, TP, FP, and FN denote the numbers of true-positive, false-positive, and false-negative, respectively.

3.3.2 3D metrics

Final assessment of the 3D detectors was based on two complementary protocols on the LUNA16 data splits defined in Section 3.1.4: (i) volumetric average precision and recall computed from ranked bounding-box detections, and (ii) free-response ROC analysis on the independent test set. Both 3D RetinaNet and the hybrid CNN-Transformer were evaluated on the same test scans and reference nodules.

Volumetric mAP and mAR. For each CT scan, the detector outputs a set of candidate volumetric bounding boxes with confidence scores. Predictions were matched to reference boxes on the same scan using three-dimensional intersection-over-union (IoU). A prediction was accepted as a true positive for threshold τ if IoU $\geq \tau$ and the reference box had not been matched previously; otherwise it was counted as a false positive. Unmatched reference boxes contributed to false negatives. Predictions from all scans in a given split were pooled and sorted globally by descending confidence before matching, following the COCO-style detection protocol extended to 3D axis-aligned boxes [13, 14, 19].

Mean average precision (mAP) was obtained by averaging AP over IoU thresholds from 0.10 to 0.50 in steps of 0.05.

For the hybrid CNN–Transformer, overlapping predictions from sliding-window inference were merged in world coordinates before mAP/mAR and FROC evaluation, so that each physical lesion contributed at most one candidate per confidence ranking step. The 3D RetinaNet used the detector's native volumetric non-maximum suppression and score filtering; both models shared the same reference boxes and patient-level splits.

Free-response ROC and mean FROC Although mAP and mAR describe detection quality based on bounding-box overlap, they do not show how sensitivity changes when more false positives per scan are allowed. This is im-

portant in lung CAD, where radiologists can review only a limited number of marks per scan [10, 12]. We therefore report free-response receiver operating characteristic (FROC) analysis on the LUNA16 test set using the official lesion-level protocol and operating points adopted in the LUNA16 challenge [10–12]. For the FROC calculation, the same post-processed candidate lists and confidence scores as for mAP/mAR were used; only the aggregation criterion differs.

The ranked list is scanned from the highest to the lowest confidence score. At each step, a candidate is evaluated with the LUNA16 lesion-matching rule (IoU ≥ 0.10). A match increments the cumulative true-positive count; a non-match increments the cumulative false-positive count.

Denoting by N_{gt} and N_{scan} the total numbers of reference nodules and test scans, and by $TP(k)$ and $FP(k)$ the cumulative true and false positives after the first k candidates, the operating sensitivity and false-positive rate are:

$$\text{Sensitivity}(k) = \frac{TP(k)}{N_{\text{gt}}}, \qquad \text{FP/scan}(k) = \frac{FP(k)}{N_{\text{scan}}}. \quad (4)$$

The FROC curve plots $\text{Sensitivity}(k)$ versus $\text{FP/scan}(k)$ as k ranges over all candidates.

Following the standard LUNA16 evaluation, we report sensitivity at seven predefined false-positive rates per scan:

$$F = \left\{\tfrac{1}{8}, \tfrac{1}{4}, \tfrac{1}{2}, 1, 2, 4, 8\right\}. \quad (5)$$

For each $f \in F$, $\text{Sensitivity@FP/scan}(f)$ is the sensitivity at the highest point on the FROC curve with $\text{FP/scan} \leq f$. The mean FROC score (mFROC), equivalent to the LUNA16 Competition Performance Metric (CPM), is the arithmetic mean of these seven values [10, 12]:

$$\text{mFROC} = \frac{1}{|F|} \sum_{f \in F} \text{Sensitivity@FP/scan}(f). \quad (6)$$

Lower FP/scan values correspond to a stricter operating point, while higher FP/scan values allow more candidate detections and therefore usually lead to higher sensitivity. Higher mFROC indicates better sensitivity across clinically relevant false-positive levels.

4. Results

This section presents the results obtained for the evaluated 2D and 3D detection models. The 2D experiments focus on the influence of annotation filtering and balancing the number of slices containing nodules and slices without nodules, whereas the 3D experiments evaluate detectors using validation metrics and test-set FROC analysis.

4.1. 2D detectors

Tables 1 - 4 summarize the test-set results obtained for all LUNA16 2D variants. For readability, the subset names are presented without the `LUNA_` prefix.

Table 1: Test-set results for RetinaNet trained on LUNA16 2D variants.

Variant	F1	Prec.	Rec.	mAP@0.5
`full`	0.5377	0.5956	0.4901	0.4036
`thr0.2`	0.6547	**0.7589**	0.5756	0.5018
`thr1.0`	0.6365	0.5317	0.7928	0.6562
`thr1.0_bal.`	0.5504	0.4011	0.8765	0.7772
`thr2.4`	0.4170	0.2731	**0.8806**	**0.8135**
`thr2.4_bal.`	**0.7024**	0.6372	0.7826	0.6971

Table 2: Additional test metrics for the 2D RetinaNet detector across LUNA16 variants.

Variant	$\text{mAP}_{0.10:0.50}$	$\text{AP}_{0.10}$	$\text{mAR}_{0.10:0.50}$	$\text{AR}_{0.10}$
`full`	0.4862	0.5276	0.5361	0.5662
`thr0.2`	0.5824	0.6083	0.6240	0.6433
`thr1.0`	0.7906	0.8417	0.8579	0.8919
`thr1.0_bal.`	**0.8936**	**0.9324**	**0.9465**	**0.9753**
`thr2.4`	0.8810	0.8863	0.9038	0.9104
`thr2.4_bal.`	0.8103	0.8437	0.8502	0.8804

For RetinaNet, progressive filtering of small annotations generally improved detection performance compared with the unfiltered full variant (see Table 1). The highest F1-score was obtained for the `thr2.4_bal.` subset (0.7024), which also achieved the best precision among all RetinaNet configurations (0.6372). In contrast, the highest mAP@0.5 value was observed for the `thr2.4` subset (0.8135), although this configuration showed substantially lower precision and F1-score.

The additional evaluation metrics presented in Table 2 showed that the `thr1.0_bal.` subset achieved the highest values of $\text{mAP}_{0.10:0.50}$ (0.8936), $\text{AP}_{0.10}$ (0.9324), $\text{mAR}_{0.10:0.50}$ (0.9465), and $\text{AR}_{0.10}$ (0.9753). The `thr2.4_bal.` configuration retained relatively high overlap-based metrics while providing a more balanced trade-off between precision and recall at the fixed operating point used for F1-score calculation. Therefore, the `thr2.4_bal.` subset was selected as the final RetinaNet-based 2D detector used in subsequent analyses.

Table 3: Test-set results for YOLO trained on LUNA16 2D variants.

Variant	F1	Prec.	Rec.	mAP@0.5
full	0.5839	0.7530	0.4768	0.5302
thr0.2	0.6555	0.8034	0.5536	0.6381
thr1.0	0.6527	0.7599	0.5721	0.6354
thr1.0_bal.	0.7421	0.7701	**0.7160**	**0.8023**
thr2.4	0.7146	**0.8586**	0.6119	0.7089
thr2.4_bal.	**0.7467**	0.8008	0.6993	0.7878

Table 4: Additional test metrics for the 2D YOLO detector across LUNA16 variants.

Variant	$mAP_{0.10:0.50}$	$AP_{0.10}$	$mAR_{0.10:0.50}$	$AR_{0.10}$
full	0.5920	0.6400	0.7704	0.8245
thr0.2	0.7298	0.7743	0.9052	0.9368
thr1.0	0.7449	0.8048	0.8949	0.9189
thr1.0_bal.	0.8327	0.8586	**0.9904**	**1.0000**
thr2.4	0.7196	0.7229	0.8458	0.8507
thr2.4_bal.	**0.8354**	**0.8652**	0.9239	0.9457

For YOLO, all filtered variants outperformed the full subset in both F1-score and mAP@0.5. The highest mAP@0.5 presented in Table 3, was achieved for the thr1.0_bal. subset (0.8023), whereas the highest F1-score was obtained for the thr2.4_bal. subset (0.7467). The thr1.0_bal. configuration simultaneously achieved high precision (0.7701) and recall (0.7160), indicating a balanced detection behavior across operating points.

The additional metrics presented in Table 4 showed that both balanced subsets achieved the highest overlap-based performance. The thr1.0_bal. subset reached the highest $mAR_{0.10:0.50}$ (0.9904) and $AR_{0.10}$ (1.0000), whereas the thr2.4_bal. subset achieved slightly higher $mAP_{0.10:0.50}$ (0.8354) and $AP_{0.10}$ (0.8652). Considering both the fixed-threshold metrics and the overlap-based evaluation, the thr1.0_bal. subset was selected as the final YOLO-based 2D detector for subsequent analyses.

Examples of predictions produced by the 2D detectors on LUNA16 test slices are presented in Fig. 1. The examples were generated using the RetinaNet and YOLO models trained on the selected LUNA16 2D dataset variants. In each image, the blue bounding box indicates the ground-truth annotation, whereas the pink bounding box marks the predicted lesion location.

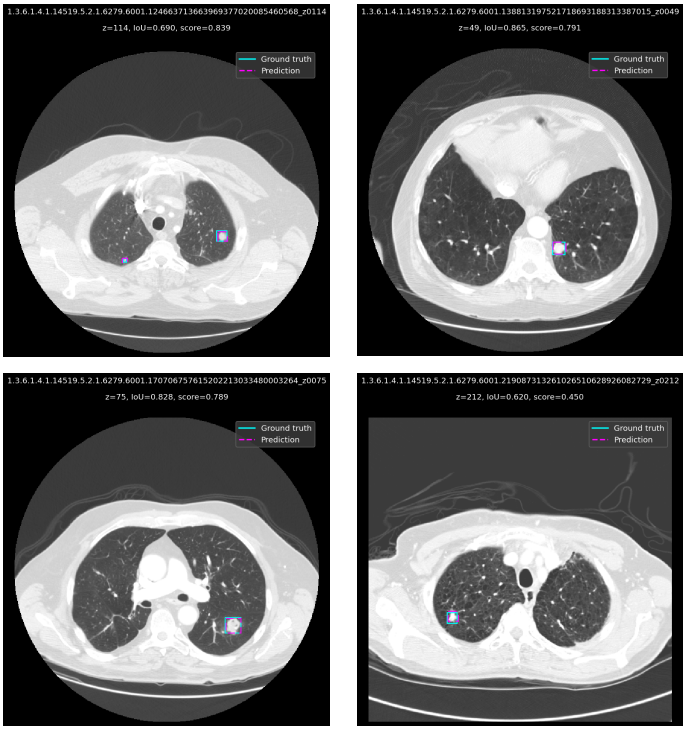

Figure 1: Example predictions of the 2D detectors on LUNA16 test slices.

4.2. 3D detectors

The main validation and test results obtained for the 3D detectors are summarized in Table 5.

RetinaNet achieved higher performance than the CNN-Transformer on both the validation and test sets. On the validation set, RetinaNet obtained mAP @ IoU 0.10-0.50 of 0.8374 and mAR of 0.9683, compared with 0.6991 and 0.8264 for the CNN-Transformer. On the test set, RetinaNet maintained high performance, reaching mAP of 0.8036 and mAR of 0.9666, whereas the CNN-Transformer achieved 0.4445 and 0.6228, respectively.

RetinaNet showed only a small decrease between validation and test performance (mAP: 0.8374 to 0.8036), while the CNN-Transformer exhibited a substantially larger drop (0.6991 to 0.4445). Similar behavior was observed for AP @ IoU 0.10 and AR @ IoU 0.10. RetinaNet maintained near-complete recovery of annotated nodules at the relaxed overlap threshold (AR @ IoU 0.10: 1.0000 on validation and 0.9910 on test), whereas the CNN-Transformer reached 0.9196 on validation and 0.7040 on test.

Table 5: Key validation and test results for the 3D detectors.

	RetinaNet		CNN-Transformer	
Metric	Validation	Test	Validation	Test
mAP @ IoU 0.10-0.50	0.8374	0.8036	0.6991	0.4445
AP @ IoU 0.10	0.8581	0.8200	0.7669	0.4980
mAR @ IoU 0.10-0.50	0.9683	0.9666	0.8264	0.6228
AR @ IoU 0.10	1.0000	0.9910	0.9196	0.7040

In addition to mAP and mAR metrics, both 3D detectors were evaluated using Free-response ROC on the independent LUNA16 test set (119 scans, 223 reference nodules), using the protocol defined in Section 3.3.2.

Table 6 reports lesion sensitivity at the seven standard FP/scan operating points. The 3D RetinaNet achieved higher sensitivity than the hybrid CNN-Transformer at every threshold. For RetinaNet, sensitivity was 0.5561 at 0.125 FP/scan and 0.8386 at 1 FP/scan, increasing to 0.9865 at 8 FP/scan; the mean FROC score (mFROC) was 0.805. For the CNN-Transformer, the corresponding values were 0.2422, 0.5381, and 0.704 at the same operating points, with mFROC 0.517. Thus, although both models produce volumetric detections on the same test data, RetinaNet offers a substantially more favorable sensitivity-false-positive trade-off under the official LUNA16 FROC criterion, whereas the CNN-Transformer reaches only moderate sensitivity even at 8 FP/scan (0.704).

Table 6: Test-set FROC sensitivity of the 3D RetinaNet detector and CNN-Transformer.

FP/scan	RetinaNet	CNN-Transformer
0.125	0.5561	0.2422
0.25	0.6457	0.3587
0.5	0.7578	0.4664
1	0.8386	0.5381
2	0.9013	0.6233
4	0.9507	0.6861
8	0.9865	0.7040
Mean	**0.8053**	**0.5170**

Examples of model predictions on CT slices from the test set are shown in Fig. 2. The predictions were obtained using the 3D RetinaNet model trained on the LUNA16 3D dataset within the MONAI framework. The blue bounding box denotes the ground-truth annotation, while the pink bounding box denotes the model prediction.

4.3. Computational cost

In addition to detection performance, inference cost was evaluated because computational efficiency is one of the main practical motivations for using 2D detectors in pulmonary nodule detection. Table 7 summarizes the mean inference time (averaged over five independent measurements) and peak GPU memory usage of the evaluated models measured on the same hardware configuration (NVIDIA GeForce RTX 3070 with 8 GB GPU memory). Since 2D detectors operate on individual axial slices, their inference time is reported per one slice. In contrast, 3D models process volumetric input and therefore their inference time is reported per CT scan; in this experiment, each scan consisted of 107 axial images.

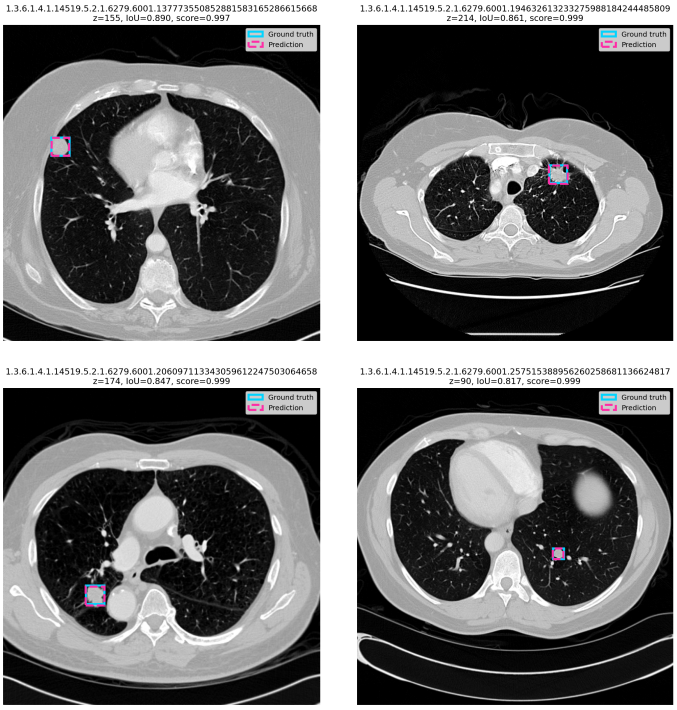

Figure 2: Example predictions of the 3D RetinaNet detector on LUNA16 test slices.

Table 7: Inference cost of the evaluated 2D and 3D detectors measured on the same hardware. Mean inference time is reported as the average of five independent runs, together with peak GPU memory usage.

Model	No of slices	Mean inference time	Peak GPU memory [MiB]
2D RetinaNet	1	54.96 ms/slice	390.68 MiB
2D YOLO	1	24.19 ms/slice	191.88 MiB
3D RetinaNet	107	12 743 ms/scan	3891.46 MiB
3D CNN-Transformer	107	340 ms/scan	1919.80 MiB

5. Conclusions

This research confirms that the choice of data representation (2D vs. 3D) and the data preparation strategy have a strong impact on the effectiveness of automatic detection of pulmonary nodules in CT images. The results clearly indicate that the volumetric 3D approach provides higher and more stable performance, primarily due to its ability to utilize the full spatial context. This is particularly important for small or partially visible lesions, which remain difficult to consistently localize on single 2D slices. Among the tested solutions, the dedicated 3D RetinaNet detector demonstrated the best generalization, achieving the highest mFROC score of 0.8053 on an independent test set and significantly outperforming the hybrid CNN-Transformer architecture in all evaluated metrics.

The conducted study also showed that 2D models can be a viable and efficient alternative when rigorous preprocessing strategies are implemented. It was shown that removing small, peripheral annotations with limited

anatomical context and class balancing significantly improved the precision and stability of lesion localization in the 2D approach. The YOLO detector optimized in this way achieved a mAP@0.5 of 0.8023, confirming that, even with limited computational or memory resources, well-prepared 2D models still offer a valuable performance-to-implementation cost ratio.

From a research and clinical perspective, these results underscore the urgent need to standardize evaluation protocols and data preparation processes. Both the design of 2D annotation variants and the selection of volumetric processing parameters significantly impact the reproducibility of experiments and the ability to compare solutions between medical centers. The advantages of 3D context should be prioritized when designing systems to support early lung cancer diagnosis, while still allowing flexibility in selecting architectures suited to the available infrastructure. The conducted experiments demonstrate the importance of direct, standardized comparisons of 2D and 3D approaches using a common FROC protocol calculated at the lesion level. In such cases, external validation of the developed solutions in real-world conditions, using datasets diverse in terms of population and hardware, is also necessary. The combination of advanced, targeted preprocessing with a carefully selected detection architecture is a key step towards reliable, clinically useful AI tools for the early detection of pulmonary nodules.

This study compared the effectiveness of 2D and 3D deep learning approaches for pulmonary nodule detection on the LUNA16 dataset, with particular focus on the influence of annotation filtering and dataset balancing in the 2D setting. The results demonstrated that preprocessing strategies applied during the creation of 2D datasets substantially affected detector performance. Removing very small peripheral annotations improved both detection quality and localization stability, while balancing the number of positive and negative slices further increased model robustness and recall.

Among the evaluated 2D configurations, the best overall performance was achieved for the balanced filtered variants. For RetinaNet, the `LUNA_thr2.4_balanced` subset obtained the highest F1-score, indicating that restricting training to larger and more clearly visible nodules improved the precision-recall trade-off. For YOLO, the `LUNA_thr1.0_balanced` configuration achieved the highest mAP@0.5 while maintaining high F1-score values. Overall, the experiments showed that 2D detectors are highly sensitive to the composition of slice-level annotations and that appropriate filtering of peripheral slices can significantly improve detection performance.

The comparison between 2D and 3D approaches showed that volumetric detectors provide substantially stronger and more stable performance for lung nodule detection. The 3D RetinaNet achieved the best overall results, outperforming the CNN-Transformer across all evaluated metrics, including mAP, mAR, and FROC sensitivity. In particular, the 3D RetinaNet maintained very high lesion sensitivity across different false-positive operating points and generalized well from validation to test data. The hybrid CNN-Transformer, although inferior to RetinaNet, also benefited from volumetric input and achieved competitive sensitivity at lower IoU thresholds, while providing more spatially consistent volumetric predictions than slice-based 2D detectors. These results indicate that incorporating full volumetric context is highly beneficial for reliable pulmonary nodule detection, especially for small or partially visible lesions that may be difficult to localize consistently in individual 2D slices.

At the same time, the promising performance obtained by the optimized 2D detectors suggests that carefully designed 2D pipelines may still represent an attractive alternative in scenarios with limited computational resources or when slice-wise processing is preferred. This observation is supported by the inference-cost analysis presented in Table 7. The selected 2D detectors required only 24.19-54.96 ms to process a single CT slice and less than 400 MiB of GPU memory, whereas the 3D RetinaNet required 12.74 s and nearly 3.9 GiB of GPU memory to process an entire CT scan. Interestingly, although the hybrid CNN-Transformer reduced the inference time to 340 ms/scan and approximately 1.9 GiB of GPU memory, this substantial improvement in computational efficiency was accompanied by markedly lower detection performance than that achieved by the 3D RetinaNet. However, the efficiency advantage of the 2D methods should be interpreted in the context of complete CT-scan analysis, since a full scan consists of 107 axial slices in the present study, whereas the 3D detectors process the entire volume jointly. Consequently, although the computational cost of the 3D RetinaNet was considerably higher, it provided the strongest lesion-level detection performance and the most robust generalization among all evaluated detectors. Nevertheless, the overall findings support the use of 3D detection frameworks as the more robust solution for CT-based lung nodule detection tasks.

6. Summary

This study investigated the performance of modern 2D and 3D deep learning approaches for pulmonary nodule detection in chest CT using the LUNA16 dataset. Multiple 2D dataset variants were prepared by filtering annotations according to bounding-box size and by balancing the proportion of positive and negative slices. The exper-

iments demonstrated that these preprocessing strategies substantially affected the effectiveness of 2D detectors.

Among the evaluated 2D configurations, the balanced filtered subsets achieved the best overall results for both RetinaNet and YOLO. The experiments also showed that removing very small peripheral annotations improved localization quality and reduced ambiguity introduced by slice-wise representation of volumetric nodules.

The comparison between 2D and 3D approaches demonstrated the advantages of volumetric analysis for pulmonary nodule detection. The 3D RetinaNet achieved the strongest overall performance, while the hybrid CNN-Transformer showed intermediate performance. This behavior may be related to the increased optimization complexity of transformer-based architectures and to the patch-based representation used by the model, which may be less effective for detecting very small nodules occupying only a limited number of voxels.

Overall, the findings support the use of 3D detection frameworks as the more robust solution for CT-based pulmonary nodule detection, while also demonstrating that appropriate preprocessing and dataset design can substantially improve the effectiveness of 2D detectors.

Acknowledgements

The research was supported in part by project "Cloud Artificial Intelligence Service Engineering (CAISE) platform to create universal and smart services for various application areas", No. KPOD.05.10-IW.10-0005/24, as part of the European IPCEI-CIS program, financed by NRRP (National Recovery and Resilience Plan) funds.

References

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2016.

[2] J. Ding, A. Li, Z. Hu, and L. Wang, "Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks," 2017.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2016.

[4] X. Wu, H. Zhang, J. Sun, S. Wang, and Y. Zhang, "Yolo-msrf for lung nodule detection," *Biomedical Signal Processing and Control*, vol. 94, p. 106318, 2024.

[5] L. Ma, G. Li, X. Feng, Q. Fan, and L. Liu, "Ticnet: Transformer in convolutional neural network for pulmonary nodule detection on ct images," *Journal of Imaging Informatics in Medicine*, vol. 37, no. 1, pp. 196–208, 2024.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[7] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sánchez, and B. van Ginneken, "Pulmonary nodule detection in ct images: False positive reduction using multi-view convolutional networks," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1160–1169, 2016.

[8] Q. Dou, H. Chen, L. Yu, J. Qin, and P.-A. Heng, "Multilevel contextual 3-d cnns for false positive reduction in pulmonary nodule detection," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1558–1567, 2017.

[9] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, E. J. R. Van Beek, D. Yankelevitz, A. M. Biancardi, P. H. Bland, M. S. Brown, R. M. Engelmann, G. E. Laderach, D. Max, R. C. Pais, D. P. Y. Qing, R. Y. Roberts, A. R. Smith, A. Starkey, P. Batra, P. Caligiuri, A. Farooqi, G. W. Gladish, C. M. Jude, R. F. Munden, I. Petkovska, L. E. Quint, L. H. Schwartz, B. Sundaram, L. E. Dodd, C. Fenimore, D. Gur, N. Petrick, J. Freymann, J. Kirby, B. Hughes, A. V. Casteele, S. Gupte, M. Sallam, M. D. Heath, M. H. Kuhn, E. Dharaiya, R. Burns, D. S. Fryd, M. Salganicoff, V. Anand, U. Shreter, S. Vastagh, B. Y. Croft, and L. P. Clarke, "Data from lidc-idri," 2015. [Data set].

[10] A. A. A. Setio, A. Traverso, T. de Bel, M. S. N. Berens, C. van den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts, *et al.*, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The luna16 challenge," *Medical Image Analysis*, vol. 42, pp. 1–13, 2017.

[11] LUNA16 Challenge, "Luna16 data." `https://luna16.grand-challenge.org/Data/`, 2016. Accessed 2026-04-29.

[12] LUNA16 Challenge, "Luna16 procedure." `https://luna16.grand-challenge.org/Procedure/`, 2016. Accessed 2026-04-29.

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

[14] M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang, *et al.*, "Monai: An open-source framework for deep learning in healthcare," *arXiv preprint arXiv:2211.02701*, 2022.

[15] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, 2017.

[16] Project MONAI, "Monai applications documentation." `https://monai-dev.readthedocs.io/en/stable/apps.html`, 2026. Accessed 2026-04-29.

[17] Project MONAI, "Monai retinanet network documentation." `https://monai-dev.readthedocs.io/en/stable/_modules/monai/apps/detection/networks/retinanet_network.html`, 2026. Accessed 2026-04-29.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*, 2014.

[20] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, 2010.