Deep Learning for Automatic Bowel-Obstruction Identification on Abdominal CT

Quentin Vanderbecq, Maxence Gelard, Jean-Christophe Pesquet, Mathilde Wagner, Lionel Arrive, Marc Zins, Emilie Chouzenoux

Deep Learning for Automatic Bowel-Obstruction Identification on Abdominal CT

Authors Names

Quentin Vanderbecq MD [a,b]

Maxence Gelard [c]

Jean-Christophe Pesquet PhD [c]

Mathilde Wagner MD, PhD[d,b]

Lionel Arrive MD, PhD[a]

Marc Zins MD[e]

Emilie Chouzenoux PhD [c]

Affiliations

[a] Department of Radiology, Saint Antoine Hospital, AP-HP.Sorbonne, 184 Rue du Faubourg Saint-Antoine, 75012 Paris, France.

[b] Université Sorbonne, Inserm U1146, CNRS UMR 7371, 15, rue de l'École de Médecine, 75006 Paris France.

[c] Université Paris-Saclay, CentraleSupélec, Inria, CVN, Gif-sur-Yvette, France.

[d] Departement of Radiology, Hospital Pitié Salpêtrière, 47-83 Bd de l'Hôpital, 75013 Paris, Île-de-France, France

[e] Departement of Radiology, Hospital Paris Saint-Joseph, 185 Rue Raymond Losserand, 75014 Paris, Île-de-France, France

***Corresponding author:** q.vanderbecq@gmail.com

Department of Radiology, Saint Antoine Hospital, AP-HP.Sorbonne, 184 Rue du Faubourg Saint-Antoine, 75012 Paris, France.

Deep Learning for Automatic Bowel-Obstruction Identification on Abdominal CT

Abstract

Rationale and Objectives: Automated evaluation of abdominal computed tomography (CT) scans should help radiologists manage their massive workloads, thereby leading to earlier diagnoses and better patient outcomes. Our objective was to develop a machine-learning model capable of reliably identifying suspected bowel obstruction (BO) on abdominal CT.

Materials and Methods: The internal dataset comprised 1345 abdominal CTs obtained in 2015–2022 from 1273 patients with suspected BO; among them, 670 were annotated as BO yes/no by an experienced abdominal radiologist. The external dataset consisted of 88 radiologist-annotated CTs. We developed a full preprocessing pipeline for abdominal CT comprising a model to locate the abdominal-pelvic region and another model to crop the 3D scan around the body. We built, trained, and tested several neural-network architectures for the binary classification (BO, yes/no) of each CT. F1 and balanced accuracy scores were computed to assess model performance.

Results: The mixed convolutional network pretrained on a Kinetics 400 dataset achieved the best results: with the internal dataset, the F1 score was 0.92, balanced accuracy 0.86, and sensitivity 0.93; with the external dataset, the corresponding values were 0.89, 0.89, and 0.89. When calibrated on sensitivity, this model produced 1.00 sensitivity, 0.84 specificity, and an F1 score of 0.88 with the internal dataset; corresponding values were 0.98, 0.76, and 0.87 with the external dataset.

Conclusion: The 3D mixed convolutional neural network developed here shows great potential for the automated binary classification (BO yes/no) of abdominal CT scans from patients with suspected BO.

Clinical Relevance Statement

The 3D mixed CNN automates bowel obstruction classification, potentially automating patient selection and CT prioritization, leading to an enhanced radiologist workflow.

List of abbreviations:

3D Mixed Convolutional network: 3DMC

Artificial intelligence: AI

Area Under Curve : AUC

Bowel Obstruction: BO

Convolutional Neural Network: CNN

Computed Tomography: CT

Long-Short-Term-Memory: LSTM

Natural Language Processing: NLP

Recurrent Neural Network: RNN

Receiver Operating Characteristics: ROC

Very Deep Convolutional Networks: VGG

Key points:

- Bowel obstruction's rising incidence strains radiologists. AI can aid urgent CT readings.

- Employed 1345 CT scans, neural networks for bowel obstruction detection, achieving high accuracy and sensitivity on external testing

- 3D mixed CNN automates CT reading prioritization effectively and speeds up bowel obstruction diagnosis.

Keywords: Abdomen. Intestines. Obstruction. Neural networks. Computed tomography.

1. INTRODUCTION

Bowel obstruction (BO) is common, with over 250 000 admissions and 2 million emergency-department visits in the US each year [1]. Moreover, annual admissions increased by 33% over the last decade [1]. BO involves the small bowel in about three-quarters of cases and the large bowel in the remaining quarter [2] (**Figure 1)**.

All guidelines recommend computed tomography (CT) as the first-line imaging study for patients with suspected BO [3, 4]. CT can confirm or refute the diagnosis of BO. When BO is present, a detailed examination of the images i) shows the site of the obstruction, notably whether the small or large bowel is involved; ii) determines whether the blockage is functional or mechanical; iii) provides details on the transition zone in case of mechanical obstruction, thereby helping to identify the cause; iv) predicts whether emergency surgery is needed, notably due to perforation or ischemia [5].

A major obstacle to the optimal use of CT in patients with suspected BO is the recent massive growth in radiologist workload. In recent years, the number of imaging studies evaluated by radiologists during on-call hours has increased severalfold [6, 7]. The advent of cross-sectional imaging has hugely increased the number of images radiologists must assess, to about one every 3–4 seconds throughout the 8-hour workday at the Mayo Clinic [8]. The constant time pressure and shorter time available for evaluating each image may increase the risk of errors [6, 9, 10]. Artificial intelligence tools can automatically select those imaging data most likely to be abnormal, which can then be given priority for a prompt and detailed radiologist reading, thereby improving patient outcomes. Such tools may be particularly valuable in emergency settings and during on-call hours [11].

The objective of this study was to identify a 3D deep-learning artificial intelligence tool capable of reliably determining whether abdominal CT scans from patients with suspected BO showed signs supporting this diagnosis.

2. MATERIAL AND METHODS

2.2 Datasets

2.2.1 Internal Dataset

We retrospectively identified all abdominal CT scans performed at a single institution between January 2015 and February 2022. We included CT scans with reports that contained the terms "obstruction" and/or "vomiting." Exclusion criteria included non-exploitable images, absence of abdominopelvic volume in CT acquisitions, patients under guardianship or trusteeship, incarcerated patients, those under legal protection, and patients objecting to the use of their data for this research. **Figure 2** is the flowchart. Three CT devices, all from the same manufacturer (GE) were used; the number of CTs acquired with each scanner is shown in **Appendix A,** and the CT acquisition technique is described in **Appendix B**. The study protocol was approved by the local institutional ethics board. A similar number of CT scans was selected at random for each year, for a total of 1500. After the exclusion of 155 unsuitable scans, 1345 CTs were left for the study.

2.2.2 Internal dataset splits

Among the 1345 CTs, we selected 620 at random, uniformly distributed over each study years. These 620 CTs were annotated by a radiologist with five years of experience in abdominal imaging. The primary criteria employed for diagnosing a small bowel obstruction includes a small bowel diameter increase of more than 25 mm [12]. Criteria for diagnosis of large BO relies on the observation of a dilated large bowel segment located proximal to a transition point, where the bowel is decompressed distal to the obstruction [13]. In cases where there is no distinct transition point, a threshold of 6 cm (8 cm in caecum) was utilized[14, 15]. When the annotation conflicted with the radiological report, the case was reviewed by at least two radiologists, including one with over ten years of experience in abdominal CT. We then selected 40 of the 620 CTs at random to serve as a test set. The

remaining 725 CTs were labeled by applying automatic natural language processing (NLP) to the radiological reports.

In the first phase of the project, only 249 CTs had been annotated, forming Dataset A. We then created Dataset B by enriching Dataset A with 250 NLP-labeled CTs to form a larger set without more input from radiologists. Then, in a second phase, as more CTs were annotated, we were able to work with an extended dataset - Dataset 1 - composed of 580 annotated CTs (not included in the test set). In the same way as in the first phase, we created Dataset 2 from Dataset 1 by adding NLP-annotated cases, this time to bring the total to 1305 CTs (all available images). As an attempt to tackle class imbalance, Dataset 3 was also created from Dataset 1 by adding 50 cases, identified as positive by NLP annotation and subsequently annotated by a radiologist (**Table 1**). Among the CTs in each dataset, 80% were assigned to the training set and 20% to the validation set. The split was performed at random and at the patient level to ensure that images from a given patient appeared in only one of the three sets.

2.3 Natural Language Preprocessing

An NLP model was developed to augment the training/validation dataset by analyzing the radiological report. First, we cleaned the text by removing special characters. Next, An open-source French language processing tool called SpaCy trained on the “fr_core_news_sm” dataset was used to simplify words into their basic forms and remove punctuation marks[16]. Additionally, we extracted any text that followed the keyword "conclusion" if it was present in the report. After this text preparation, we applied a NLP model based on term frequency-inverse document frequency (TF-IDF) was then applied to create a numerical representation of the text, and logistic regression was used to build a predictive model. The NLP was trained on 228 reports. On the test dataset comprising 40 radiologist-annotated CTs, the F1 score of

the resulting model was 0.88. The model was used to label the 725 CT reports that were not annotated by the radiologist.

2.4 CT Preprocessing

CT provides high-resolution 3D images. Volumes in the CT database are represented as 3D arrays, with 300 to 500 slices, each containing 512x512 pixels. We first applied a clipping procedure to the Hounsfield units, setting the level at 50 and using a window of 350, to help normalize the intensity values of the CT images. We then extracted the abdominal region using the ResNet-18 model pretrained on the ImageNet dataset (https://www.image-net.org) [17]. We added a decision-making component to this model to label the areas as 'thorax,' 'abdomen,' or 'pelvis. After training on 4425 axial slices from eight CTs, model accuracy was 0.93. The label for each axial slice was deduced, and the location of the cut on the CT scan was determined by identifying the intersection between the curves corresponding to the "abdomen" and "thorax" categories. The Simple Linear Iterative Clustering (SLIC) method was used with superpixels to crop the image with minimal void around the body [18]. This previous step of the preprocessing procedure produced CT volumes of average size (369, 484, 369), which were rescaled to (256, 224, 224) using bilinear interpolation. Voxel intensity was normalized to achieve a mean of 0 and a standard deviation of 1 to maintain consistent intensity levels across the CT scans.

2.5 Models

We assessed the following six, supervised, deep-learning models for binary classification of CT scans as BO yes/no. (i) The 3D Mixed Convolutional network (3DMC), an 18-layer mixed convolutional network [19], was pretrained on Kinectics 400 dataset [20] (**Appendix C**). (ii) The fully 3D convolutional neural network Res3D, an 18-layer ResNet

(ResNet-18) 3D network [19], was pretrained on a video dataset (Kinectics 400 dataset [20]). (iii) The Convolutional Neural Network (CNN)/Long-Short-Term-Memory (LSTM) with a 2D ResNet-18 network, pretrained on ImageNet, as the CNN and a Recurrent Neural Network (RNN) to capture the 3D nature of CT. Each axial slice was processed by ResNet-18, and the resulting sequence of feature representations was handled as a time series and passed through the LSTM unit (**Appendix D**). (iv) The big CNN/LSTM model differed from the previous CNN/LSTM model in that it used a deeper ResNet, with 34 layers (ResNet-34), instead of 18. (v) The VGG/LSTM, also exhibiting a similar architecture to the two previous models, used a very deep convolutional network (VGG) instead of ResNet-18 [21]. (vi) Finally, the 2.5D CNN/LSTM, based on the classic CNN/LSTM architecture, used ResNet2D instead of ResNet3D, but with a stack of five 2D slices instead of a single 2D slice as model input, thus producing a small 3D volumetric representation (hence the term "2.5D").

To address database imbalance, we used a balanced cross-entropy loss function, which gives more importance to positively classified samples during training (**Appendix E**).

2.6 External test set

We retrospectively studied abdominal CT scans performed between March 23, 2022, and May 29, 2022, at an institution different from the one that provided the internal dataset and where the CT machines were from a different manufacturer (Siemens Healthcare). Out of the 905 CT reports indicating that abdominal images were acquired, 88 contained the terms "obstruction" and/or "vomiting". All 88 CTs were annotated using the same protocol as for the internal dataset, by a radiologist with five years of experience in abdominal CT. Again, when the annotation conflicted with the radiological report, the case was reviewed by at least two radiologists, including one with over ten years of experience in abdominal CT.

2.7 Classification performance

To compare the six models, we computed balanced accuracy as the sum of sensitivity and specificity divided by 2, and the F1 score as 2 × Precision × Recall/(Precision + Recall). The receiver operating characteristics (ROC) curve was plotted to calibrate the model on sensitivity. The computations were performed on the training and validation sets using the PyTorch framework, while the test set results were processed using the scikit-learn Python library. Area Under the Curve (AUC) of ROC 95% confidence intervals were generated using 5,000 bootstrap samples.

3. RESULTS

3.1 Patients characteristics

Within the institution used to construct internal datasets, a total of 10,389 CT scans met the inclusion criteria. Additionally, there were 134 CTs from patients who declined to participate in the study. An equal number of CT scans were randomly selected for each year, resulting in a total of 1,500 scans. After excluding 155 unsuitable scans, we were left with 1,345 CT scans for the study. These 1345 CTs of Internal dataset were obtained from 1273 patients (55% female; median age, 69 [23–92] years). **Table 2** displays patients characteristics.

For the external dataset, 88 CTs were obtained from 86 patients (56% female; median age, 65 [22–94] years).

3.2 Binary classification performance of the six models

We focused initially on Dataset 1 composed only of radiologist-annotated CTs. Based on a grid search for parameters in the CNN/LSTM and Res3D models, we selected the following parameter settings: learning rate equal to 0.001; Stochastic Gradient Descent

optimizer; weight decay equal to 0.1; and training over 40 epochs. All six models were optimized on the training set. Then, the validation set was used to compare the models and to set parameters such as the learning rate, weight decay, and criteria for stopping the training stage. Finally, the test set was used for the final assessment of model performance.

The F1 score and balanced accuracy of each model were as follows: 3D Mixed Convolutional (3DMC), **0.89** and **0.86**; Res3D, 0.87 and 0.84; CNN/LSTM (with ResNet-18), 0.69 and 0.61; big CNN/LSTM (with ResNet-34), 0.68and 0.60; VGG/LSTM, 0.69 and 0.61; and 2.5D CNN/LSTM, 0.62 and 0.63. An analysis of the loss training and validation curves for the LSTM models showed that performance improved on the training dataset, but reached a plateau, then started to decay on the validation set, suggesting overfitting on the training data and difficulties in generalizing effectively to previously unseen data. All the LSTM-based models appeared to be limited in terms of generalization capabilities **(Figure 3)**. **Figure 4a** confirms the above ranking. In particular, it shows that it is not dependent on the threshold set at 0.5 on the binary classification task, as it remains consistent overall the ROC curves, i.e., for any threshold value.

The best performance was obtained with the 3D Mixed Convolutional Network (3DMC), which was therefore used for the rest of the study.

3.3 Performance of the 3DMC on the different datasets

We initially evaluated the 3DMC model on Datasets 1 and 2, both containing the same 580 annotated CTs. We assessed the change in performance induced by adding data with NLP labelling. Performance was best on the smaller dataset (Dataset 1), with F1 scores of 0.89 and 0.82 on the training and validation sets, respectively (**Figure 5**).

With Dataset 3, despite the smaller imbalance, performance was not better on the internal test set: the highest F1 score was 0.86 compared to 0.89 with Dataset 1. However, the

training curve exhibited a more stable behaviour (**Figure 5**). Similarly, as demonstrated by Figure 4a, **Figure 4b** demonstrated that the ranking is not contingent on the binary decision threshold.

3.4 Hyperparameter tuning

We evaluated two learning rates (0.001 and 0.0006) and four weight decays (0.5, 0.1, 0.01, and 0.001). The model performance appeared to be robust with respect to the choice of the learning rate or weight decay (**Appendix F**). We then used these parameters to stop training early, with the goal of optimizing the loss or the F1 score on the validation dataset. We obtained three models for subsequent evaluation on the test set.

The first of these three models was trained on Dataset 1 with an early stopping rule designed to optimize the F1 score. The second model was also trained on dataset 1, with an early stopping rule to optimize the loss curve. The third model was trained on dataset 3 and optimized F1 score and Loss. The results are reported in **Table 3.**

3.5 Model calibration on sensitivity

Our objective was to maximize sensitivity in detecting CT scans with evidence of BO to ensure that no patient with BO would be missed. **Figure 6** shows the ROC curve for the validation set, considering all possible thresholds for the softmax outputs of the optimized 3DMC. The highest sensitivity was associated with insufficient specificity. We therefore selected the second-highest sensitivity (0.98), which was associated with the highest specificity. On the internal test set, this model had 1.00 sensitivity, 0.84 specificity, and an F1 score of 0.88. On the external test set, corresponding values were 0.98, 0.76, and 0.87. A single CT was incorrectly identified as having no evidence of BO; upon radiologist review, it showed cecal volvulus with a high position of the gas-distended cecum.

3.6 Classification performance on small datasets

We trained the 3DMC on the two smaller datasets: Dataset A (249 radiologist-annotated CTs) and Dataset B (Dataset A plus 250 NLP labeled CTs). The settings were as follows: learning rate equal to 0.001; stochastic gradient descent; weight decay equal to 0.1; and 40 epochs. Contrary to the finding with Datasets 1 and 2, performance improved when NLP-positive CTs were added (**Appendix G**).

3.7 Timing of Inference

Towards the conclusion, we constructed a pipeline that fuses all preprocessing steps from DICOM files, and we gauged the time required to achieve a classification inference on our machine, utilizing a single GPU. In the case of the external dataset, we achieved results with a median inference time per file of about 29 seconds (range: 2-31 seconds).

4. DISCUSSION

We developed a machine-learning model specifically designed to identify abdominal CT scans that support a previously clinically suspected diagnosis of bowel obstruction (BO). The automated CT classification into two groups, with and without evidence of the suspected diagnosis, enables radiologists to prioritize CT scans most likely to indicate BO and promptly identify patients who require emergency treatment. On both internal and external datasets, our model achieved remarkable performance with 100% and 98% sensitivity, respectively, for identifying CTs that showed evidence of BO, as confirmed by radiologist interpretation. These sensitivity values are comparable to those obtained by radiologists [12, 22]. Thus, the high sensitivity of our model ensures that it effectively identifies cases that require urgent

attention. Consequently, the resulting time reduction to deliver appropriate treatment is expected to lead to improved patient outcomes [23, 24].

To develop our model, we created a preprocessing pipeline that was not previously available for abdominal CT. This pipeline directly displayed the model predictions for CT scans in DICOM format. Performance and reproducibility of our model were excellent on an external dataset acquired using CT machines from a different manufacturer than for the internal dataset. Notably, sensitivity was 0.98 on the external dataset. This performance was comparable to that of a radiologist [12, 22], within the context of comparing clinical-surgical and radiological gold standard.

Despite its high prevalence, particularly in emergency departments [1], digestive obstruction has received limited attention in research. Two studies conducted by the same team investigated the use of standard radiography for diagnosis [25], although this approach is no longer recommended for managing digestive obstruction [3, 4]. One paper [26] focusing on identifying the localization of transition zones, yielded mixed results. Our model focuses on the identification of bowel obstruction without encompassing the entirety of the radiologist's analysis. Building on the insights from this study, our proposal was to revisit the algorithm for the diagnostic management of BO, starting from its initial stages. These foundational steps, besides serving as a triage tool, would also furnish valuable data for training more sophisticated models and address other questions encountered by radiologists when interpreting CT scans for BO. One such question includes determining the location of the transition zone in cases of mechanical occlusion. By reassessing and refining the diagnostic approach and leveraging this information to develop advanced models, we aim to enhance the overall accuracy and efficiency of diagnosing digestive obstruction, especially in emergency settings.

Coupling a deep-learning binary-classification model with 2D and 3D convolutions produced the best results. The LSTM architecture was initially developed for video classification [19] but in our study, it performed below our expectations. Contrary to our initial hypothesis, which was that LSTMs would closely mimic radiologists' interpretation by scrolling through 2D images to capture the 3D nature of CT scans, they did not perform as well as anticipated. Similarly, Multiscale Vision Transformer pretrained on color videos did not perform well on our datasets (data not shown) [27]. These observations may be related to the fact that such architectures would require additional training data to reach good performance. The used models are primarily designed for color images, whereas most medical images are in grayscale. To address this limitation, we merged the RGB kernels in the first layer of all our network architectures, allowing effective processing of grayscale images.

A key finding from our study was that adding NLP-labeled CTs significantly improved model performance when the datasets were small. Results reported with data augmentation techniques such as rotations and cropping [28] led us to expect that increasing dataset size by adding NLP-labeled CTs would improve performance. This was not consistently the case: performance improved when the number of CTs increased from 249 (annotated) to 499 (annotated+NLP-labeled), but not from 580 (annotated) to 1305 (annotated+NLP-labeled). The first assumption behind the failure of data augmentation is the limited performance of the NLP model itself. However, it is essential to acknowledge that achieving 100% agreement with an LP model is likely not feasible, especially considering that some cases may require radiologist consensus. While NLP can offer valuable additional labels, it is critical to ensure that the NLP model performance is reliable enough. However, as the dataset continues to expand, the reannotation process by experts will inevitably become more time-consuming. Nevertheless, obtaining a substantial amount of high-quality labeled data directly from experts in the field remains crucial. By carefully managing the

contributions of NLP-based labels and direct annotations, we can enhance the model performance and ensure its effectiveness in the classification task.

One limitation of our study is that all annotations were performed by a single radiologist, possibly introducing some level of subjectivity and error. However, all disagreements between this radiologist and the radiological reports were examined by several expert radiologists to achieve a consensus. Second, the training data originated from a single center, where all the CT machines were from the same manufacturer. This point might limit the applicability of our findings to CTs acquired using machines from other manufacturers. However, the results proved reproducible in the external dataset, in which the CTs were obtained using machines from a different manufacturer. Third, we do not know whether the model performed as well as radiologists' reproducibility. Our research has revealed that only earlier studies, which are somewhat dated, have examined the consistency of radiologists' diagnoses of bowel obstruction during CT scan evaluations. One study reported a kappa value of 0.92 [29]. Fourth, we did not distinguish between small-bowel and large-bowel obstructions The radiologist directly categorizes the examination into three classes: no obstruction, small bowel obstruction, and colonic obstruction. Exploring classification into three categories, namely, without BO, with BO of the small bowel, and with BO of the large bowel may constitute a promising research perspective. Fifthly, one limitation in comparing our models arises from the absence of cross-validation, as it was not carried out due to significant computational costs, given that a model demands approximately 2 days of training. Rather than relying on statistical comparisons of k-fold, we opted for alternative approaches, such as external outcomes.

Finally, it is important to highlight that our study is not specifically designed to assess the clinical impact of the model. A comprehensive evaluation in a clinical routine setting will require further investigation. Evaluating how the model accelerates radiology workflow can

be a challenging task. It is crucial to emphasize that our model's primary focus is not to assist radiologists in making diagnoses but to streamline and smooth the workflow by prioritizing potentially pathological scans. This objective is in line with the typical application of AI models in emergency radiology [11, 30]. Furthermore, with the advancement of teleradiology, where a single radiologist can remotely interpret images from multiple centers, AI has the potential to improve the prioritization of image interpretation, especially when multiple studies arrive simultaneously. While we did not explicitly evaluate this aspect in our study, our model's strong performance in terms of sensitivity may contribute to reducing errors and guiding radiologists' attention toward likely pathological cases.

In conclusion, when applied to CTs obtained to assess suspected BO, a model combining supervised 2D and 3D convolutional neural networks effectively identified those CTs showing evidence of BO. Leveraging this model has the potential to enhance the triage process for abdominal scans and therefore to appropriate patient management, notably during on-call hours. Furthermore, this approach acts as the first step in automating the selection of patients with mechanical small BO and assessing the severity of CT signs that indicate the need for surgical management.

References

1. Peery AF, Crockett SD, Murphy CC, et al (2019) Burden and Cost of Gastrointestinal, Liver, and Pancreatic Diseases in the United States: Update 2018. Gastroenterology 156:254-272.e11. https://doi.org/10.1053/j.gastro.2018.08.063
2. Johnson WR, Hawkins AT (2021) Large Bowel Obstruction. Clin Colon Rectal Surg 34:233–241. https://doi.org/10.1055/s-0041-1729927
3. ten Broek RPG, Krielen P, Di Saverio S, et al (2018) Bologna guidelines for diagnosis and management of adhesive small bowel obstruction (ASBO): 2017 update of the evidence-based guidelines from the world society of emergency surgery ASBO working group. World Journal of Emergency Surgery 13:24. https://doi.org/10.1186/s13017-018-0185-2
4. Expert Panel on Gastrointestinal Imaging, Chang KJ, Marin D, et al (2020) ACR Appropriateness Criteria® Suspected Small-Bowel Obstruction. J Am Coll Radiol 17:S305–S314. https://doi.org/10.1016/j.jacr.2020.01.025
5. Zins M, Millet I, Taourel P (2020) Adhesive Small Bowel Obstruction: Predictive Radiology to Improve Patient Management. Radiology 296:480–492. https://doi.org/10.1148/radiol.2020192234
6. Bruls RJM, Kwee RM (2020) Workload for radiologists during on-call hours: dramatic increase in the past 15 years. Insights into Imaging 11:121. https://doi.org/10.1186/s13244-020-00925-z
7. Lantsman CD, Barash Y, Klang E, et al (2022) Trend in radiologist workload compared to number of admissions in the emergency department. European Journal of Radiology 149:. https://doi.org/10.1016/j.ejrad.2022.110195
8. McDonald RJ, Schwartz KM, Eckel LJ, et al (2015) The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. Acad Radiol 22:1191–1198. https://doi.org/10.1016/j.acra.2015.05.007
9. Hames K, Patlas MN, Mellnick VM, Katz DS (2019) Errors in Emergency and Trauma Radiology: General Principles. In: Patlas MN, Katz DS, Scaglione M (eds) Errors in Emergency and Trauma Radiology. Springer International Publishing, Cham, pp 1–16
10. Patel AG, Pizzitola VJ, Johnson CD, et al (2020) Radiologists Make More Errors Interpreting Off-Hours Body CT Studies during Overnight Assignments as Compared with Daytime Assignments. Radiology 297:374–379. https://doi.org/10.1148/radiol.2020201558
11. Jalal S, Parker W, Ferguson D, Nicolaou S (2021) Exploring the Role of Artificial Intelligence in an Emergency and Trauma Radiology Department. Can Assoc Radiol J 72:167–174. https://doi.org/10.1177/0846537120918338
12. Fukuya T, Hawes DR, Lu CC, et al (1992) CT diagnosis of small-bowel obstruction: efficacy in 60 patients. American Journal of Roentgenology 158:765–769. https://doi.org/10.2214/ajr.158.4.1546591
13. Jaffe T, Thompson WM (2015) Large-Bowel Obstruction in the Adult: Classic Radiographic and CT Findings, Etiology, and Mimics. Radiology 275:651–663. https://doi.org/10.1148/radiol.2015140916
14. Taourel P, Kessler N, Lesnik A, et al (2003) Helical CT of large bowel obstruction. Abdom Imaging 28:267–275. https://doi.org/10.1007/s00261-002-0038-y
15. Khurana B, Ledbetter S, McTavish J, et al (2002) Bowel Obstruction Revealed by Multidetector CT. American Journal of Roentgenology 178:1139–1144. https://doi.org/10.2214/ajr.178.5.1781139
16. Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020) spaCy: Industrial-strength Natural Language Processing in Python.
17. He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition.

In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp 770–778
18. Achanta R, Shaji A, Smith K, et al (2010) SLIC superpixels. Technical report, EPFL
19. Tran D, Wang H, Torresani L, et al (2018) A Closer Look at Spatiotemporal Convolutions for Action Recognition
20. Kay W, Carreira J, Simonyan K, et al (2017) The Kinetics Human Action Video Dataset
21. Simonyan K, Zisserman A (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition
22. Furukawa A, Yamasaki M, Furuichi K, et al (2001) Helical CT in the diagnosis of small bowel obstruction. Radiographics 21:341–355. https://doi.org/10.1148/radiographics.21.2.g01mr05341
23. Perotte R, Lewin GO, Tambe U, et al (2018) Improving Emergency Department Flow: Reducing Turnaround Time for Emergent CT Scans. AMIA Annual Symposium Proceedings 2018:897
24. Wang DC, Parry CR, Feldman M, et al (2015) Acute Abdomen in the Emergency Department: Is CT a Time-Limiting Factor? American Journal of Roentgenology 205:1222–1229. https://doi.org/10.2214/AJR.14.14057
25. Cheng PM, Tejura TK, Tran KN, Whang G (2018) Detection of high-grade small bowel obstruction on conventional radiography with convolutional neural networks. Abdom Radiol (NY) 43:1120–1127. https://doi.org/10.1007/s00261-017-1294-1
26. Vanderbecq Q, Ardon R, De Reviers A, et al (2022) Adhesion-related small bowel obstruction: deep learning for automatic transition-zone detection by CT. Insights Imaging 13:13. https://doi.org/10.1186/s13244-021-01150-y
27. Fan H, Xiong B, Mangalam K, et al (2021) Multiscale Vision Transformers
28. Perez L, Wang J (2017) The Effectiveness of Data Augmentation in Image Classification using Deep Learning
29. Maglinte DD, Gage SN, Harmon BH, et al (1993) Obstruction of the small intestine: accuracy and role of CT in diagnosis. Radiology 188:61–64. https://doi.org/10.1148/radiology.188.1.8511318
30. Weisberg EM, Chu LC, Fishman EK (2020) The first use of artificial intelligence (AI) in the ER: triage not diagnosis. Emerg Radiol 27:361–366. https://doi.org/10.1007/s10140-020-01773-6

TABLES

Table 1: Summary of data distribution within datasets

Dataset	Total n of CTs	Radiologist-annotated CTs	NLP-labeled CTs	Radiologist-annotated NLP-positive CTs
Dataset A	249	249	0	0
Dataset B	499	249	250	0
Dataset 1	580	580	0	0
Dataset 2	1305	580	725	0
Dataset 3	630	580	0	50

n: number; CTs: computed tomography scans; NLP: natural language processing (via artificial intelligence)

Table 2: Patients characteristics

	Internal Set (n=670*)	External Set (n=88)
CLINICAL		
Sex (%)		
Women	362 (54)	49 (56)
Men	308 (46)	39 (44)
Age (years)	69 [23–92]	65 [22–94]
Clinical Context		
Emergency	387 (58)	61 (69)
Inpatient	212 (32)	26 (30)
Outpatient	71 (10)	1 (1)
OBSTRUCTION	215 (32)	47 (53)
Small bowel	167 (78)	34 (72)
Large bowel	44 (20)	12 (26)
Others	4 (2)	1 (2)
Etiologies		
Single Adhesive Band	61 (28)	20 (43)
Tumor	29 (13)	3 (6)
Hernia	19 (9)	3 (6)
Others	35 (16)	6 (13)
Volvulus	12 (6)	5 (11)
Carcinomatosis	11 (5)	3 (6)

Matted Adhesions	14 (7)	2 (4)
Ileitis	8 (4)	3 (6)
Stomia	6 (3)	0
Fecaloma	14 (7)	1 (2)
Ischemic	6 (3)	1 (2)

Continuous variable is displayed in median [range] and categorical variables in number (percentage). *Patients from radiologist annotated data

Table 3: Optimization of the 3D mixed convolutional network selected as the best model[a]

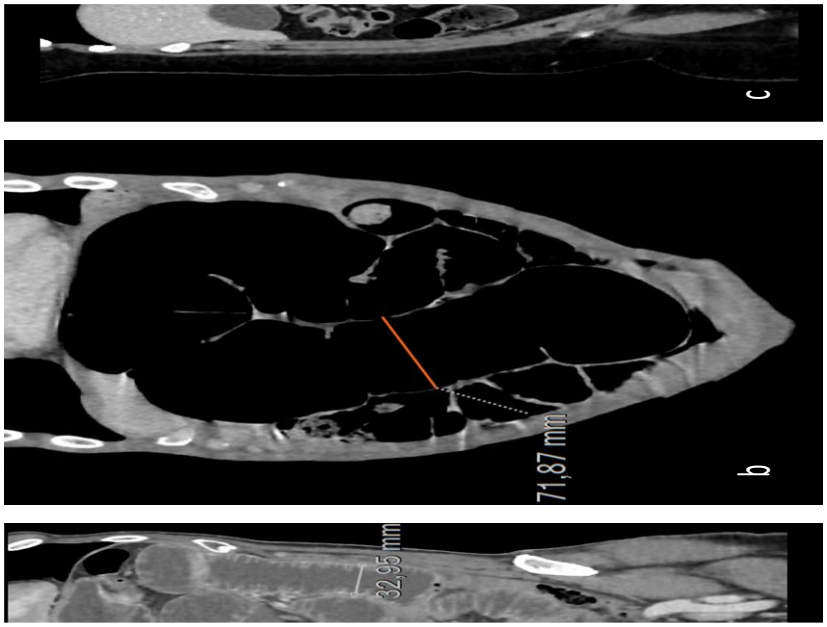
	F1 score optimized on Dataset 1	***Loss optimized on Dataset 1***	***F1 score and Loss simultaneously optimized on Dataset 3***
Internal test set	**F1**: 0.94	**F1**: 0.92	F1: 0.92
	Bal.Acc.: 0.88	**Bal.Acc**.: 0.86	**Bal.Acc**.: 0.86
	Se: 0.95	**Se**: 0.93	**Se**: 0.93
External test set	**F1**: 0.85	**F1**: 0.89	**F1**: 0.89
	Bal.Acc.: 0.86	**Bal.Acc**.: 0.89	**Bal.Acc**.: 0.89
	Se: 0.85	**Se**: 0.89	**Se**: 0.89

[a]An early stopping rule was used on the validation set to optimize the F1 score and/or loss and the resulting optimized model was evaluated on an internal test set and an external test set.

F1: F1 score; Bal.Acc.: balanced accuracy; Se: sensitivity (recall)

FIGURES

Figure 1: Computed tomography scans of the abdomen and pelvis

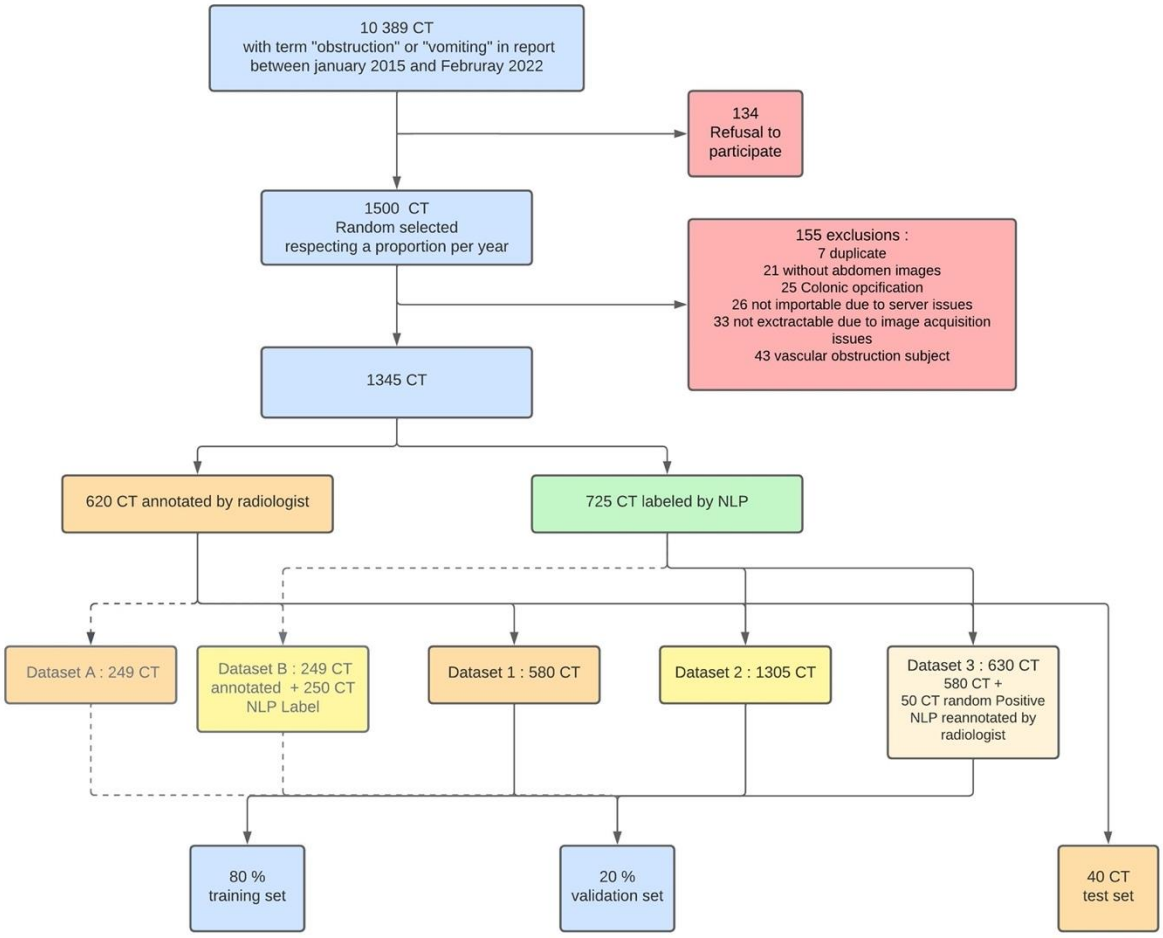


a. Small bowel obstruction: a loop has a diameter greater than 25 mm (32.95 mm)

b. Large bowel obstruction: a loop has a diameter greater than 60 mm (71.87 mm)

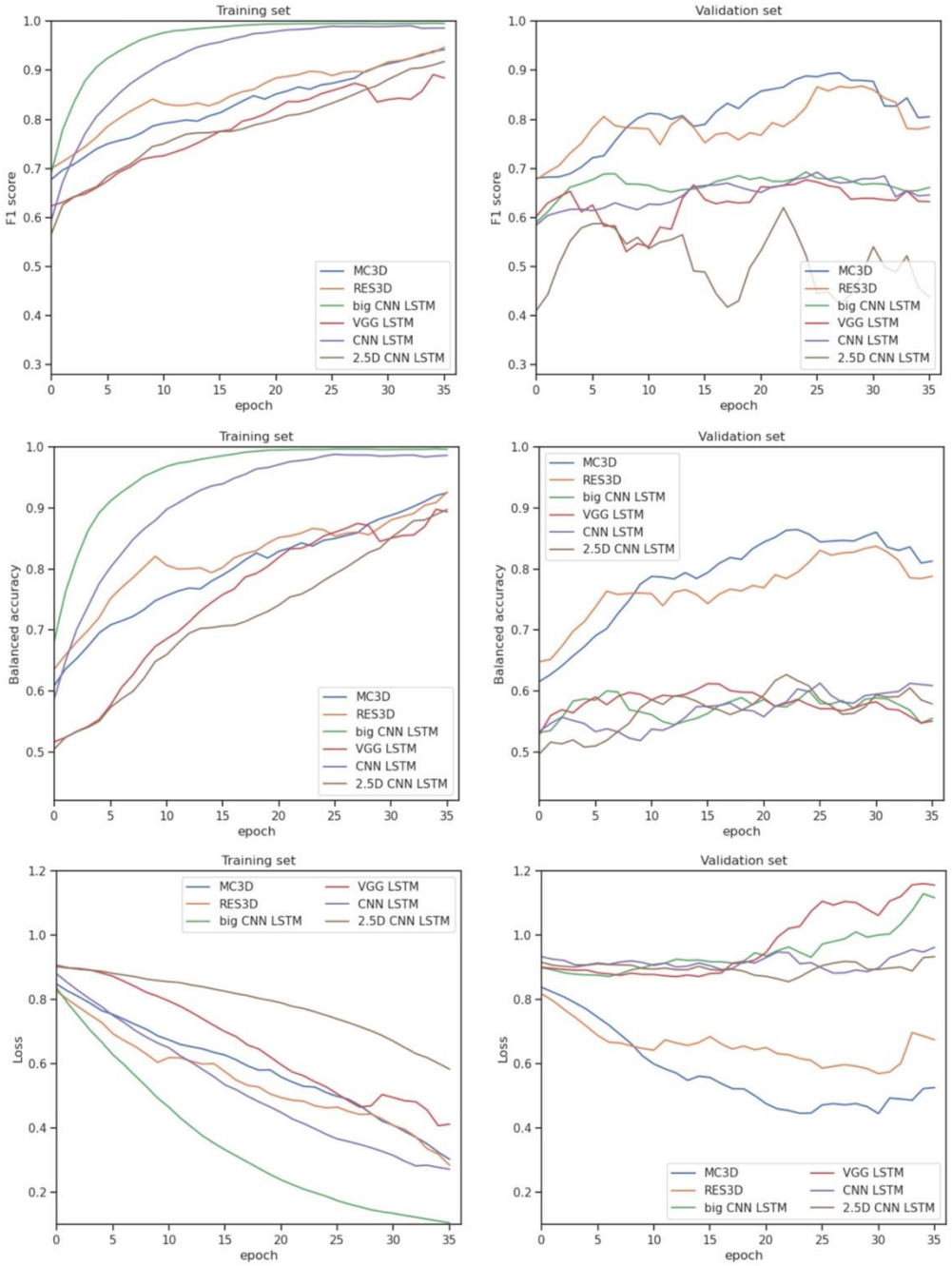
c. Normal findings (no bowel dilation)

Figure 2: Flowchart of the data selection process



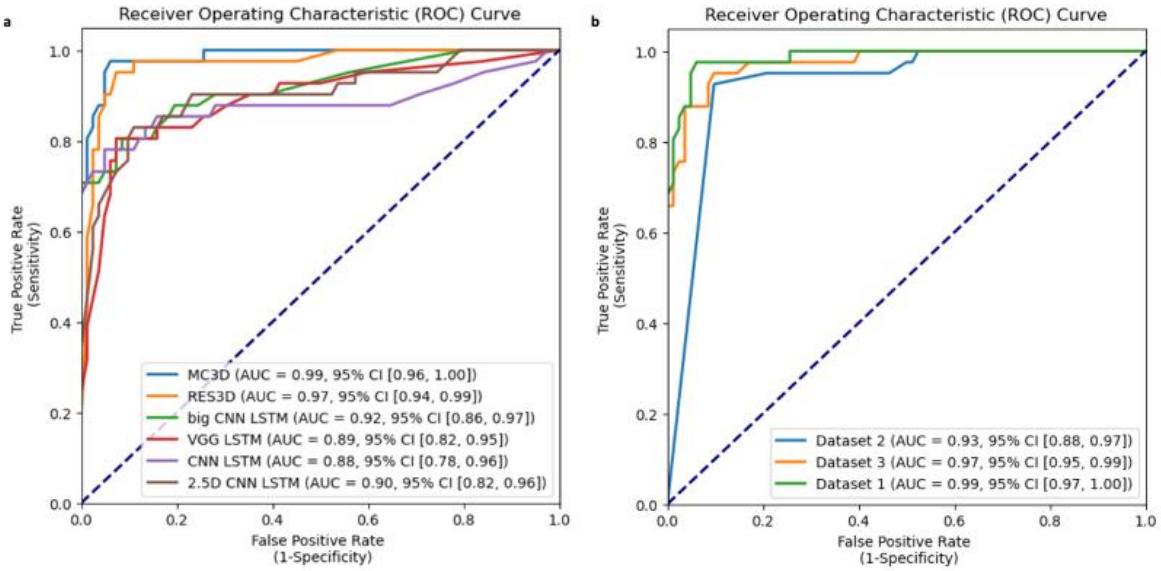
CT: computed tomography scan; NLP: Natural Language Processing

Figure 3: Compared performance of six models



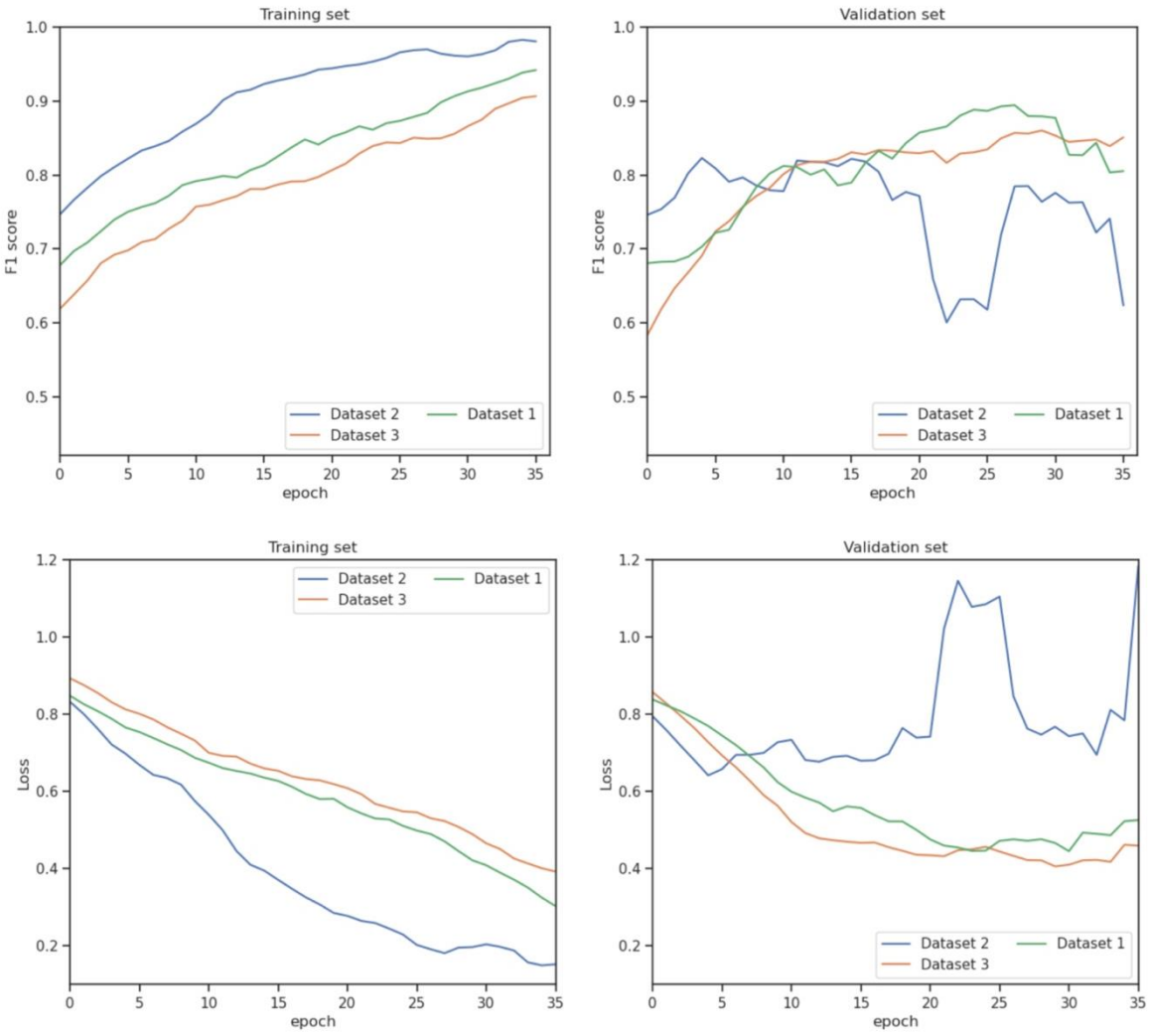
MC3D: 3D mixed convolutional network; RES 3D: 3D ResNet pretrained on a video dataset; big CNN LSTM: big convolutional neural network-recurrent neural network with ResNet-34 as the CNN; VGG LSTM: very deep convolutional network-recurrent neural network; CNN LSTM: convolutional neural network-recurrent neural network, with ResNet-18 as the CNN; 2.5D CNN LSTM: convolutional neural network-recurrent neural network with a stack of five 2D slices as the model input

Figure 4: ROC Curves Comparing the Performance of differents Models (a) and the 3D Mixed Convolutional Model across the three Datasets (b) on internal test set



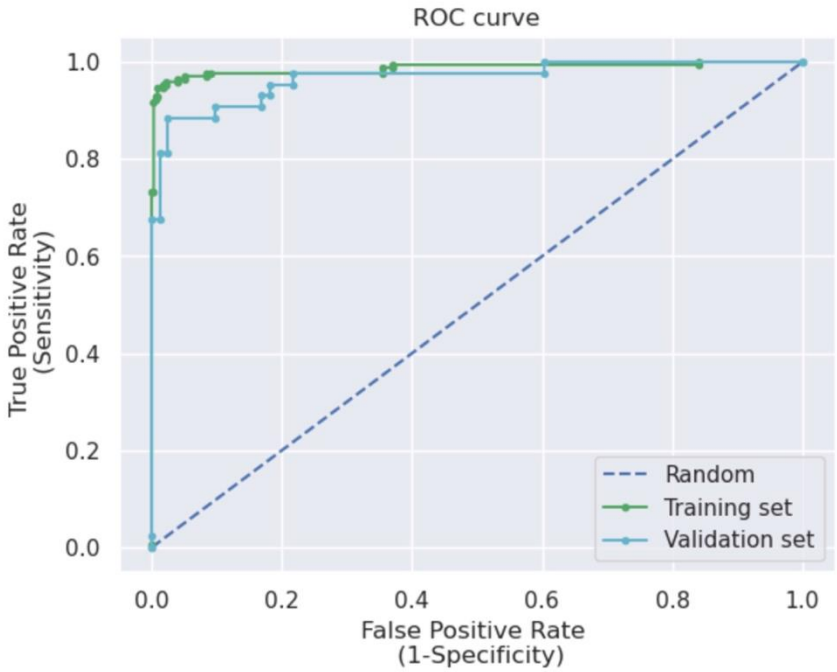
MC3D: 3D mixed convolutional network; RES 3D: 3D ResNet pretrained on a video dataset; big CNN LSTM: big convolutional neural network-recurrent neural network with ResNet-34 as the CNN; VGG LSTM: very deep convolutional network-recurrent neural network; CNN LSTM: convolutional neural network-recurrent neural network, with ResNet-18 as the CNN; 2.5D CNN LSTM: convolutional neural network-recurrent neural network with a stack of five 2D slices as the model input. Dataset 1 was composed of 580 CTs annotated by a radiologist. Dataset 2 comprised 1305 CTs including the 580 CTs in Dataset 1. Dataset 3 combined the 580 annotated CTs and 50 CTs labeled as positive by Natural Language Processing then annotated by a radiologist.

Figure 5: Performance of the selected model on different datasets



Dataset 1 was composed of 580 CTs annotated by a radiologist. Dataset 2 comprised 1305 CTs including the 580 CTs in Dataset 1. Dataset 3 combined the 580 annotated CTs and 50 CTs labeled as positive by Natural Language Processing then annotated by a radiologist.

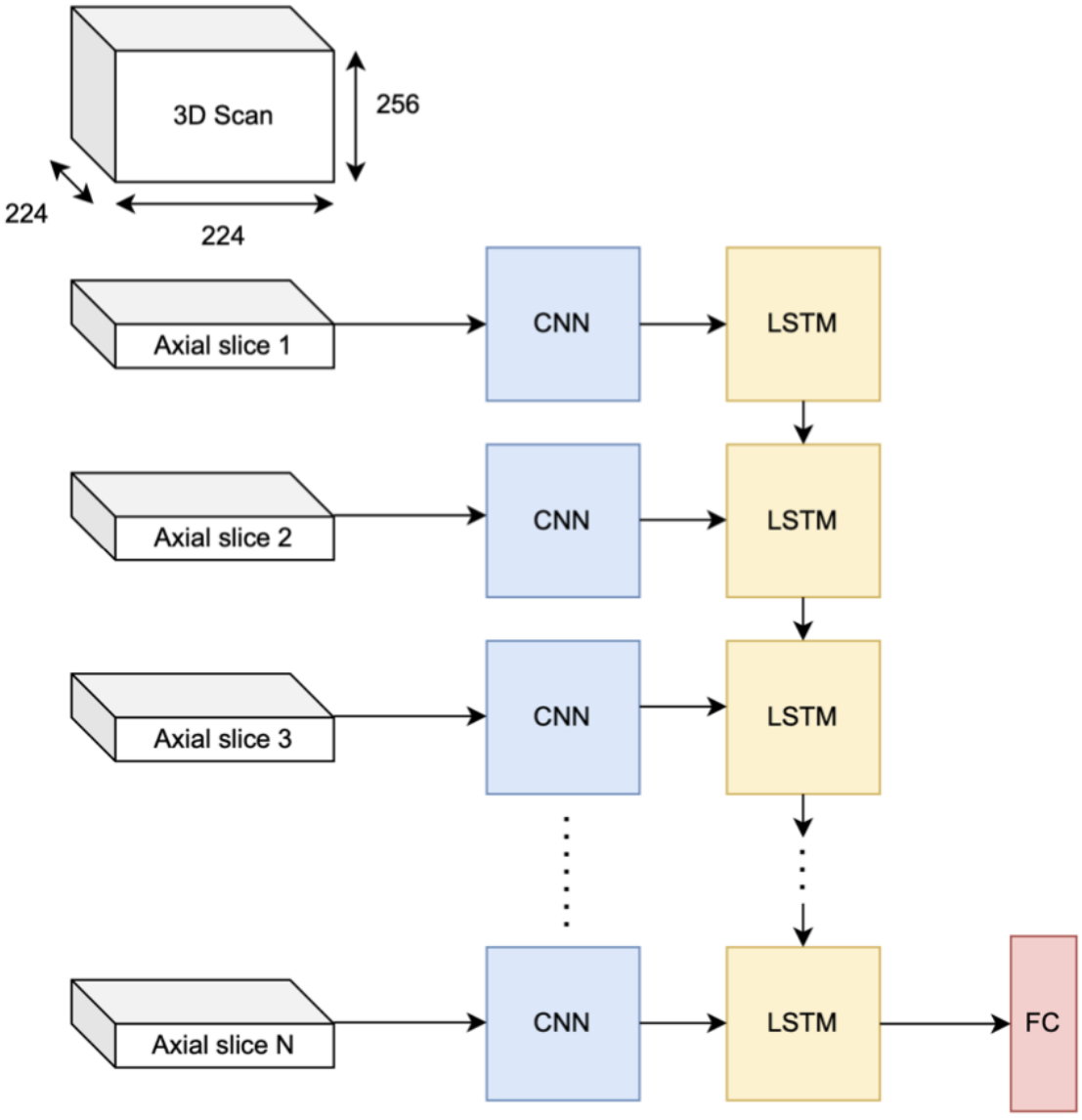
Figure 6. Receiver Operating Characteristic curve for the selected model (3D mixed convolutional network)



All thresholds of the final softmax output from the optimized 3D mixed convolutional network were evaluated.

APPENDICES

Appendix A: LSTM model architecture



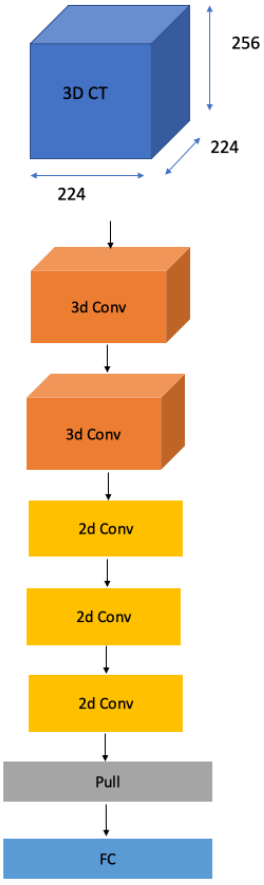
CNN: convolutional neural network (ResNet-18, ResNet-34, or VGG); LSTM: long-short-term-memory; FC: fully connected

Appendix B: Distribution across datasets of computed tomography scans showing bowel obstruction

Bowel obstruction	No, n (%)	Yes, n (%)
Internal Test Set, n=	32 (80)	8 (20)
Training and Validation Sets, annotated CTs from Dataset 1, n=	423 (67)	207 (33)
External Test Set, n=	41 (47)	47 (53)

mm, reconstruction section thickness of 1.25 mm, and 1.25-mm gap. Oral contrast material was not used in any of the patients.

Appendix C: 3D Mixed Convolutional network architecture



CT: computed tomography. Conv: convolutional layer. FC: fully connected

Appendix D: Number of radiologist-annotated computed tomography (CT) scans obtained with each of the three CT machine models in each internal dataset

CT model	Training/Validation Set	Test Set
Revolution™ CT[a]	32	2
Revolution™ EVO[a]	220	17
Revolution™ Frontier[a]	378	41

[a]GE Healthcare, Chicago, IL

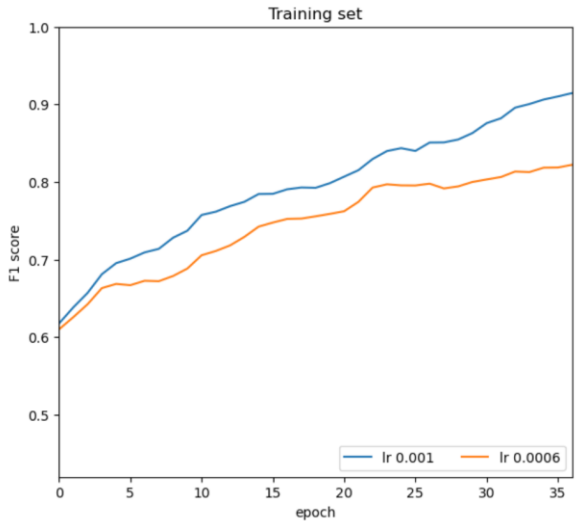
Appendix E: CT acquisition

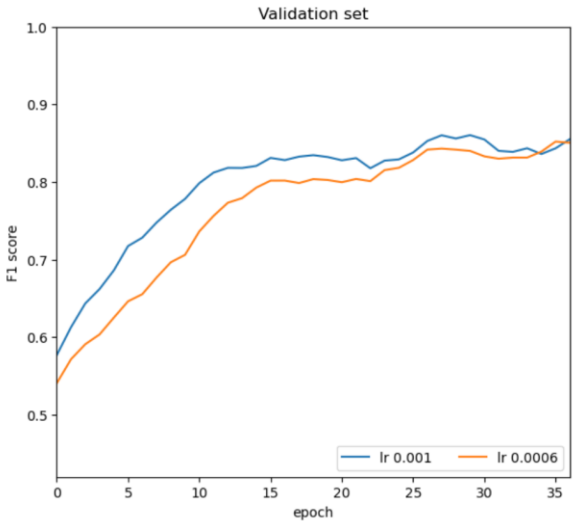
All CT images were obtained using 64-detector machines (Appendix A), with 1.375 pitch, 0.7 tube rotation time, and 120 kV. Images were first acquired without contrast material and with a nominal section thickness of 0.625 mm, reconstruction section thickness of 1.25 mm, and 1.25-mm gap. Then, an iodinated contrast agent (iopromide, iohexor, or iomeprol) was administered intravenously in a dose of 1.5 mL/kg and at a flow rate of 3 mL/s. Images were acquired at the portal phase (70 s after the injection), with a section thickness of 0.625

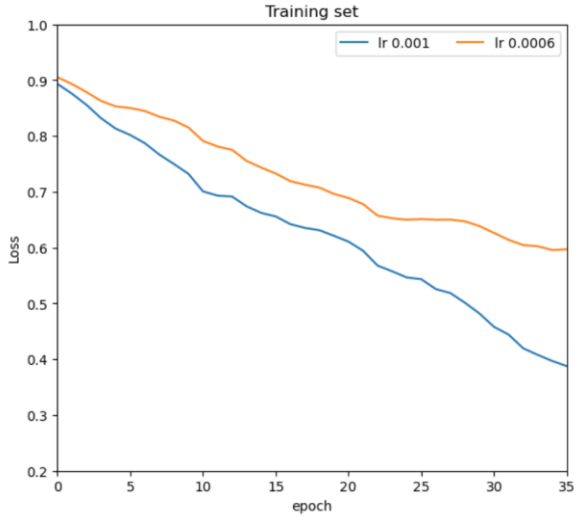
Injection	Training/Validation Set	Test Set
No	59	2
Yes	571	38

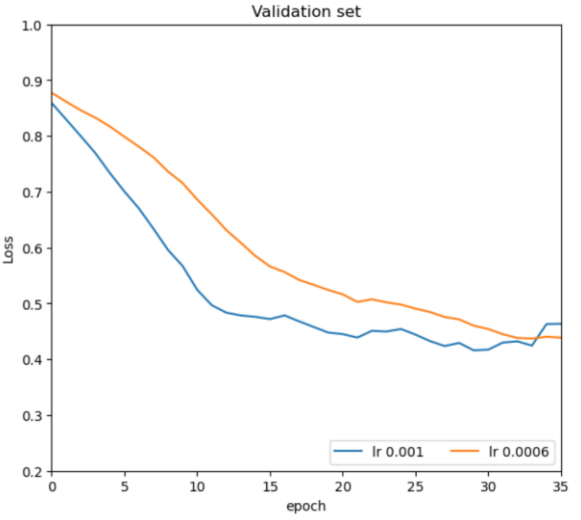
Number of computed tomography (CT) scans injected at portal phase on our database

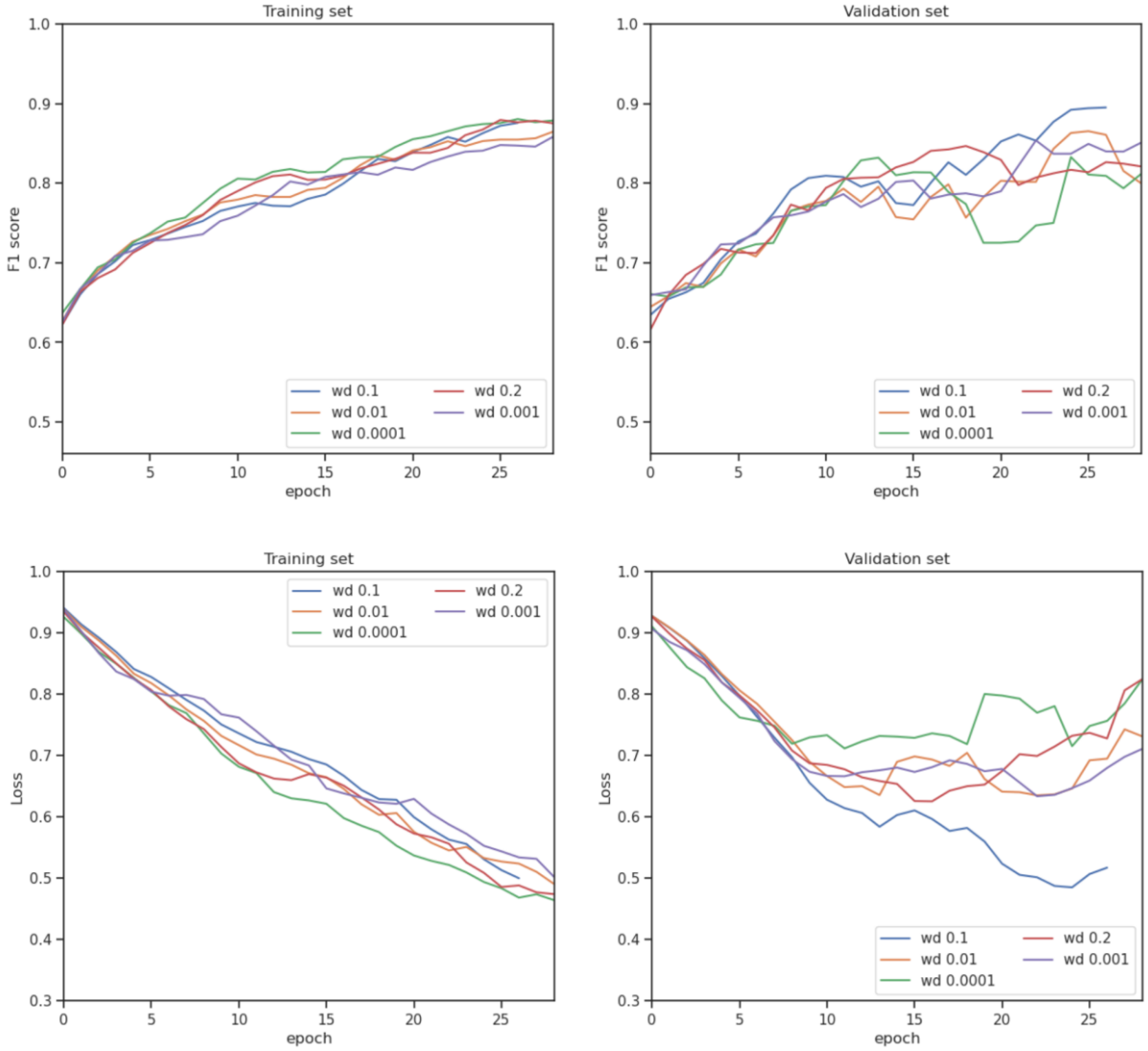
Appendix F: Performance of the selected model with different learning rate and weight decay values





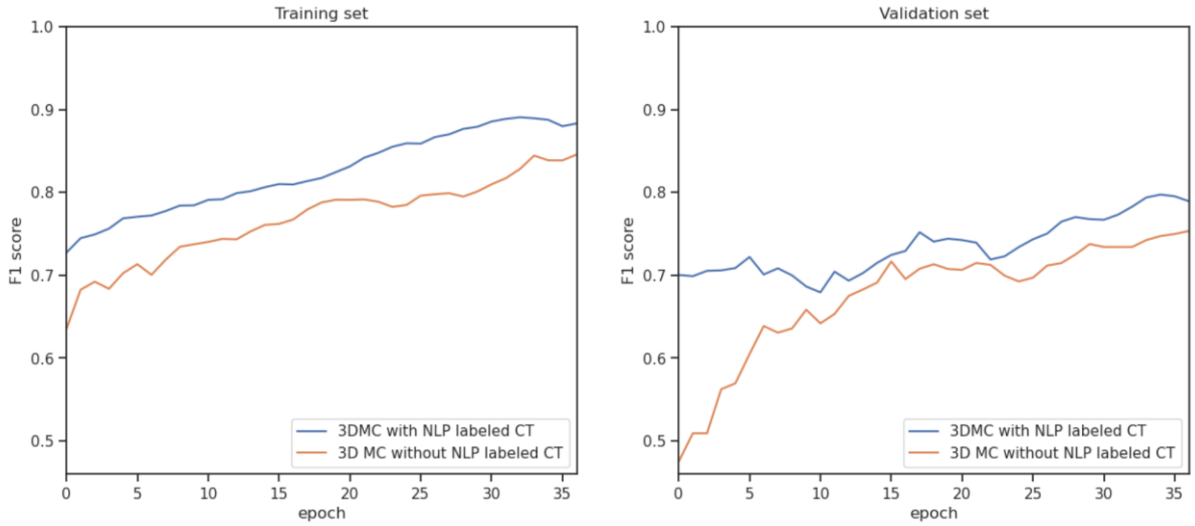






lr: learning rate; wd: weight decay

Appendix G: Performance of the selected model (3D mixed convolutional network, 3DMC) with and without the addition of NLP-labeled CTs to a small dataset



The model was trained and validated using 249 CTs annotated by a radiologist (orange curves) and with these same 249 CTs combined with 250 CTs labeled using natural language processing (NLP).