Title

Generalizability of Deep Learning Classification of Spinal Osteoporotic Compression Fractures on Radiographs Using an Adaptation of the Modified-2 Algorithm-Based Qualitative Criteria

Authors

Qifei Dong MS

Department of Biomedical Informatics and Medical Education, University of Washington

Seattle, WA 98195, USA

Gang Luo PhD

Department of Biomedical Informatics and Medical Education, University of Washington

Seattle, WA 98195, USA

Nancy E. Lane MD

Department of Medicine, University of California - Davis

Sacramento, CA 95817, USA

Li-Yung Lui MA MS

Research Institute, California Pacific Medical Center

San Francisco, CA 94143, USA

Lynn M. Marshall, ScD

Epidemiology Programs, Oregon Health and Science University-Portland State University School of Public Health

Portland, OR 97239, USA

Sandra K. Johnston, PhD, RN

Department of Radiology, University of Washington

Seattle, WA 98195-7115, USA

Howard Dabbous MD

Department of Radiology and Imaging Sciences, Emory University

Atlanta, GA 30322, USA

Michael O'Reilly MBBCh MSc MPH

Department of Radiology, University of Limerick Hospital Group

Limerick, Ireland

Ken F. Linnau MD MS

Department of Radiology, University of Washington

Seattle, WA 98195-7115, USA

Jessica Perry MS

Department of Biostatistics, University of Washington

Seattle, WA 98195, USA

Brian C. Chang MD MScr

Department of Biomedical Informatics and Medical Education, University of Washington

Seattle, WA 98195, USA

Jonathan Renslo MS

Keck School of Medicine, University of Southern California

Los Angeles, CA 90033, USA

David Haynor MD PhD

Department of Radiology, University of Washington

Seattle, WA 98195-7115, USA

Jeffrey G. Jarvik MD MPH

Departments of Radiology and Neurological Surgery, University of Washington

Seattle, WA 98104-2499, USA

Nathan M. Cross MD MS

Department of Radiology, University of Washington

1959 NE Pacific Street

Box 357115

Seattle, WA 98195-7115, USA

https://orcid.org/0000-0002-5040-4340

Originating Institution

University of Washington

1959 NE Pacific Street

Box 357115

Seattle, WA 98195-7115, USA

Corresponding Author

Nathan M. Cross MD MS

Department of Radiology, University of Washington

1959 NE Pacific Street

Box 357115

Seattle, WA 98195-7115, USA

Email: nmcross@uw.edu

Phone: 206 598-2870

Acknowledgments

Research and results reported in this publication were partially facilitated by the generous contribution of computational resources from the Department of Radiology of the University of Washington.

Funding

This research was 1) supported by the University of Washington Clinical Learning, Evidence, And Research (CLEAR) Center for Musculoskeletal Disorders, Administrative, Methodologic and Cores and NIAMS/NIH grant P30AR072572; and 2) supported in part by the General Electric-Association of University Radiologists Radiology Research Academic Fellowship (GERRAF), a career development award co-sponsored by General Electric Healthcare and the Association of University Radiologists. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

The Osteoporotic Fractures in Men (MrOS) Study is supported by National Institutes of Health funding. The following institutes provide support: the National Institute on Aging (NIA), the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS), the National Center for Advancing Translational Sciences (NCATS), and NIH Roadmap for Medical Research under the following grant numbers: U01 AG027810, U01 AG042124, U01 AG042139, U01 AG042140, U01 AG042143, U01 AG042145, U01 AG042168, U01 AR066160, R01 AG066671, and UL1 TR000128.

Gang Luo was partially supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under Award R01HL142503. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests

No author that analyzed or controlled the data used in this research is employed by or consulted by any company in the medical industry.

Mr. Dong reports grants from NIH/NIAMS during the conduct of the study.

Dr. Luo reports grants from NIH/NIAMS and grants from NIH/NHLBI during the conduct of the study.

Ms. Lui reports grants from NIH/NIA during the conduct of the study.

Dr. Johnston reports grants from NIH/NIAMS during the conduct of the study.

Dr. Chang reports grants from NLM during the conduct of the study.

Dr. Haynor reports grants from NIH/NIAMS during the conduct of the study.

Dr. Jarvik reports grants from NIH/NIAMS during the conduct of the study; and Springer Publishing: Royalties as a book co-editor; GE Healthcare for the GE-Association of University Radiologists Radiology Research Academic Fellowship (GERRAF): Travel reimbursement for Faculty Board of Review; Wolters Kluwer/UpToDate: Royalties as a chapter author.

Dr. Cross reports grants from NIH/NIAMS during the conduct of the study.

Abstract

Rationale and Objectives: Spinal osteoporotic compression fractures (OCFs) can be an early biomarker for osteoporosis but are often subtle, incidental, and under-reported. To ensure early diagnosis and treatment of osteoporosis, we aimed to build a deep learning vertebral body classifier for OCFs as a critical component of our future automated opportunistic screening tool.

Materials and Methods: We retrospectively assembled a local dataset including 1,790 subjects and 15,050 vertebral bodies (thoracic and lumbar). Each vertebral body was annotated using an adaption of the modified-2 algorithm-based qualitative criteria. The Osteoporotic Fractures in Men (MrOS) Study dataset provided thoracic and lumbar spine radiographs of 5,994 men from six clinical centers. Using both datasets, five deep learning algorithms were trained to classify each individual vertebral body of the spine radiographs. Classification performance was compared for these models using multiple metrics including the area under the receiver operating characteristic curve (AUC-ROC), sensitivity, specificity, and positive predictive value (PPV).

Results: Our best model, built with ensemble averaging, achieved an AUC-ROC of 0.948 and 0.936 on the local dataset's test set and the MrOS dataset's test set, respectively. After setting the cutoff threshold to prioritize PPV, this model achieved a sensitivity of 54.5% and 47.8%, a specificity of 99.7% and 99.6%, and a PPV of 89.8% and 94.8%.

Conclusion: Our model achieved an AUC-ROC >0.90 on both datasets. This testing shows some generalizability to real world clinical datasets and a suitable performance for a future opportunistic osteoporosis screening tool.

Keywords

Osteoporosis, osteoporotic fracture, deep learning, opportunistic screening, radiography

Main Body

INTRODUCTION

Osteoporosis affects 9% of individuals over 50 years old in the US (1) and 200 million women globally (2). In developed countries, one out of three individuals will suffer an osteoporotic compression fracture (OCF) in their lifetime (2). After the first OCF, the risk for subsequent OCFs increases greatly (3-5). Even one OCF can decrease quality of life and increase risk of mortality (6).

Osteoporosis screening is evidence-based and is endorsed by many organizations, including the US Preventive Services Task Force, but remains underutilized. Between 2004 and 2006, more than 2/3 of women who should have been screened for osteoporosis were not (7). From 2006-2010, screening of US women with Medicare using dual-energy X-ray absorptiometry decreased by 56% (8). The rate of osteoporosis screening for high-risk men is also low (9).

Opportunistic osteoporosis screening, which uses pre-existing imaging to increase osteoporosis detection rates, can complement current osteoporosis screening methods and is desired to introduce minimum extra cost. Several approaches to opportunistic osteoporosis screening have been proposed (10-29). Many research groups used computerized tomography (CT) images (10-22), while few used radiographs (23-29). Radiography is a ubiquitous imaging modality used early in diagnostic workup of many conditions with an estimated 183 million exams in US hospitals in 2010 (30). Thus, using radiographs to conduct opportunistic osteoporosis screening is as important as using CT and could potentially reach a broader patient population. Using radiographs, Lee et al. (23) and Zhang et al. (24) used machine learning algorithms to estimate bone mineral density. However, using bone mineral density as a biomarker of osteoporosis detection has known limitations (31, 32). Spinal OCFs can serve as an additional osteoporosis biomarker and are often incidental on chest or abdominal images and frequently under-reported, resulting in under-diagnosis and under-treatment (33). Applying automated opportunistic OCF screening to existing imaging studies could result in earlier and more extensive osteoporosis identification and treatment. Multiple studies (25-29) have attempted to automatically detect OCFs using

radiographs. However, these studies had limitations including single center data leading to possible overfitting (25-28) and unclear dataset construction processes (29).

We ultimately aim to build an automated opportunistic OCF radiograph screening tool with three primary sequential components (see Figure 1). Adequate performance of any clinical test can only be judged in the context of the use case. Considering a screening tool for large volumes of studies, a tool with too many false positives could unduly burden the health care system. Thus, we prioritized positive predictive value (PPV) and specificity of the model rather than sensitivity.

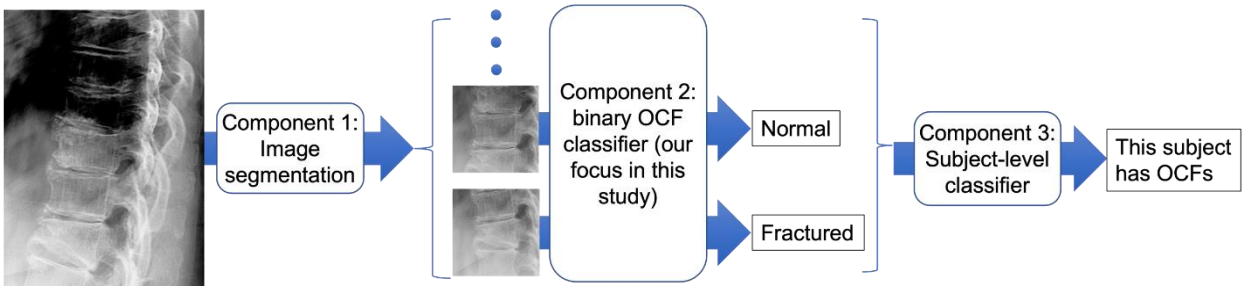


Figure 1. Our future automated opportunistic screening tool detecting OCFs on radiographs. This tool has three components: 1) image segmentation and extraction of vertebral bodies; 2) a binary OCF classifier predicting whether each vertebral body has a moderate to severe OCF or not; and 3) a subject-level classifier integrating the OCF predictions of all vertebral bodies with additional structured data to determine this subject's OCF status.

In this paper, we focus on the second component, the binary OCF classifier (see Figure 1). This component predicts whether an image patch containing a single vertebral body (termed vertebral patch) has a moderate to severe OCF or not. The first component, which is used to automatically extract the individual vertebral patches, is a distinct body of work (34). To develop the OCF classifier in this study, we extracted each vertebral body using manually annotated corner points.

The current work in this paper is an extension of the work in (35), in which spine radiographs from the Osteoporotic Fractures in Men (MrOS) Study (36, 37) were used. In the current work, we used two spine

radiograph datasets with multicenter data: 1) a dataset assembled from multiple clinical sites across a single local healthcare enterprise (hereafter termed the local dataset) and 2) the MrOS dataset. These two datasets include only thoracic and lumbar spine radiographs because OCFs are rare in the rest of the axial skeleton. To detect OCF on each vertebral patch, we used deep learning, the state-of-the-art technique for image classification. Our objective is to train a performant and generalizable OCF classifier with an area under the precision-recall (PR) curve (AUC-PR) >0.70 and an area under the receiver operating characteristic (ROC) curve (AUC-ROC) >0.90 on the multicenter data mentioned above.

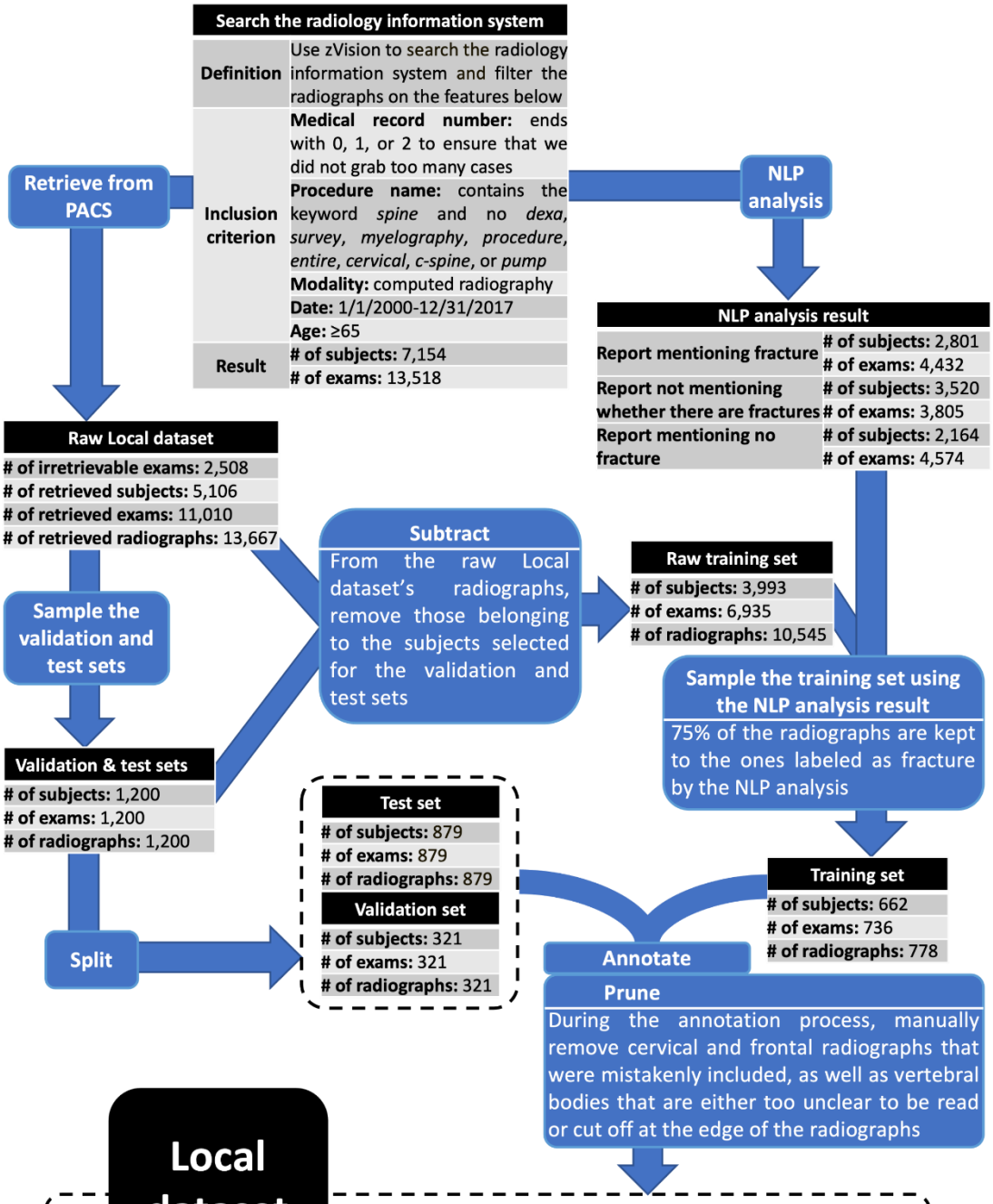
MATERIALS AND METHODS

Brief introduction to the datasets

We obtained two datasets containing lateral thoracic and lumbar spine radiographs: the clinically derived local dataset and the research MrOS dataset (36, 37). The local dataset contains clinical data for diagnostic purposes, while the MrOS dataset was generated for research. To make the deep learning models performant on clinical data, we typically used the local dataset to fine-tune the models. Both datasets were used to test the models.

Local dataset

This dataset contains clinical data acquired in varied clinical settings for diagnostic purposes. The spine radiographs in this retrospective dataset were acquired from 2000 to 2017 at multiple clinical sites (inpatient, outpatient, and emergency) across a single healthcare enterprise. The mean ages (±standard deviation) of female and male subjects were 75±8 years and 75±9 years, respectively. Figure 2 shows the construction of this dataset.



Local dataset

Test set	Validation set	Training set
# of subjects: 847	# of subjects: 307	# of subjects: 636
# of exams: 847	# of exams: 307	# of exams: 706
# of radiographs: 847	# of radiographs: 307	# of radiographs: 742

# of vertebral bodies		Label 0			Label 1		
		Normal	Non-osteoporotic deformity	Vertebral body's height loss <20%	Vertebral body's height loss in 20%-25%	Vertebral body's height loss in 26%-40%	Vertebral body's height loss >40%
Training set	Thoracic	2,221	243	142	75	84	80
	Lumbar	2,081	501	106	115	110	105
Validation set	Thoracic	1,014	37	36	26	12	11
	Lumbar	1,070	94	28	24	14	15
Test set	Thoracic	3,050	61	111	37	38	69
	Lumbar	2,896	167	68	51	55	47

Figure 2. Construction of the local dataset and partitioning it into the training, validation, and test sets. zVision (Intelerad; Montreal, Canada), a radiology information search tool, queried the radiology information system (RIS) to identify subjects and exams fitting the inclusion criteria. A natural language processing (NLP) system called LireNLPSystem (41) analyzed each exam's radiology report to roughly determine whether it described a fracture. The NLP result for each exam's radiology note served as a weak label for this exam. These weak labels could help roughly balance the training set. Radiographs of the subjects that satisfied the inclusion criteria were retrieved from the picture archiving and communication system (PACS). From the retrieved radiographs, we randomly selected 1,200 subjects and a single radiograph of each subject to form the validation and test sets. From these 1,200 radiographs, 879 were randomly assigned to form the test set and the remaining 321 were assigned to the validation set. To avoid overlap between the training set and the other two sets, the radiographs in the training set were sampled from the 13,667 radiographs excluding those of the 1,200 subjects that had been selected for the validation and test sets. To form the training set, 778 radiographs were sampled. To improve the balance of the training set, 75% of the radiographs were randomly sampled from the ones labeled as 'fracture' by NLP. The remaining radiographs were randomly sampled from the ones labeled as 'no fracture' or 'not mentioned' by NLP. Finally, the local dataset was annotated. Further data pre-processing and augmentation steps (including other data balancing steps) are introduced in Section A of the Supplemental Materials.

Two of the co-authors reviewed each radiograph to guarantee that they were de-identified and contained no protected health information. All radiographs were originally in the Digital Imaging and Communications in Medicine (DICOM) format. The DICOM tags, which could contain protected health information, were removed by converting the DICOM radiograph to Tag Image File Format.

On each radiograph in the local dataset, we annotated each vertebral body's four corner points and severity of OCF using DicomAnnotator (38), an open-source annotation software. Multiple groups participated in the process of annotating the corner points of each vertebral body. The OCF severity of

each vertebral body was annotated using the modified-2 algorithm-based qualitative (m2ABQ) criteria (39), a revised version of the modified algorithm-based qualitative (mABQ) criteria (40). Five individuals annotated OCF severity of each vertebral body, including three faculty radiologists (27, 17, and 10 years of experience, respectively), one neuroradiology fellow (7 years of experience), and one biomedical informatics graduate student. This process consisted of 17 rounds. We randomly split the local dataset into 17 subsets. In the first eight rounds, at least two individuals annotated each radiograph. For each of these first eight rounds, we computed Fleiss' kappa and Cohen's kappa to measure the inter-reader agreement, and held a consensus meeting to discuss the disputed annotations. In the last nine rounds, each radiograph was annotated by one annotator. More details about the local dataset annotation are presented elsewhere (39).

Classification systems and radiologists struggle to accurately classify mild or subtle OCFs often confounded by parallax artifact, remote traumatic injuries, and congenital variations (40). Our future opportunistic screening tool is intended to complement the current clinical standard of care while introducing a minimum of extra cost. Including mild OCFs into our classification system could substantially increase false positives, which would cause more downstream cost. Our use case, to alert or not alert a provider or radiologist to a potentially missed fracture, required a binary classification, defined as highly probable OCF vs normal/non-osteoporotic deformity/mild or questionable fracture. Therefore, we dichotomized the m2ABQ categories: "label 0" representing normal/non-osteoporotic deformity/mild or questionable fracture vs. "label 1" representing moderate or severe fracture.

The local dataset was partitioned into the training, validation, and test sets. As shown in Figure 2, the training set was balanced for better model training. In contrast, we kept the class distributions of the validation and test sets consistent with those in the original population.

MrOS dataset

The de-identified MrOS dataset was obtained from the San Francisco Coordinating Center under a data use agreement. This dataset was generated for research and includes only male subjects, and thus has lower diversity than the local dataset. Details (including population information) for the MrOS dataset are presented in multiple papers (35-37, 42). Six US academic medical centers (36, 37) contributed data to this dataset.

The MrOS team had previously annotated the MrOS dataset based on a modification (42) of the Genant semiquantitative (mSQ) criteria (43). To determine OCFs, the mSQ criteria require the presence of endplate depression, making these criteria closer to the mABQ criteria (40). To adapt to our binary OCF classification, the mSQ categories were simplified into two classes (moderate or severe fracture vs. normal/trace/mild fracture) (35). This is similar to the m2ABQ simplification previously discussed.

From the MrOS dataset's test set, we randomly selected 122 radiographs containing 844 vertebral bodies, each assigned an m2ABQ label. Table 1 shows the number of vertebral bodies for each (dataset, OCF classification criteria) combination. In the rest of this paper, each of these combinations is denoted by "dataset-classification criteria." For example, MrOS-m2ABQ denotes the dataset whose data are from the MrOS dataset and are annotated using the m2ABQ criteria.

Table 1. The number of vertebral bodies for each (dataset, OCF classification criteria) combination.

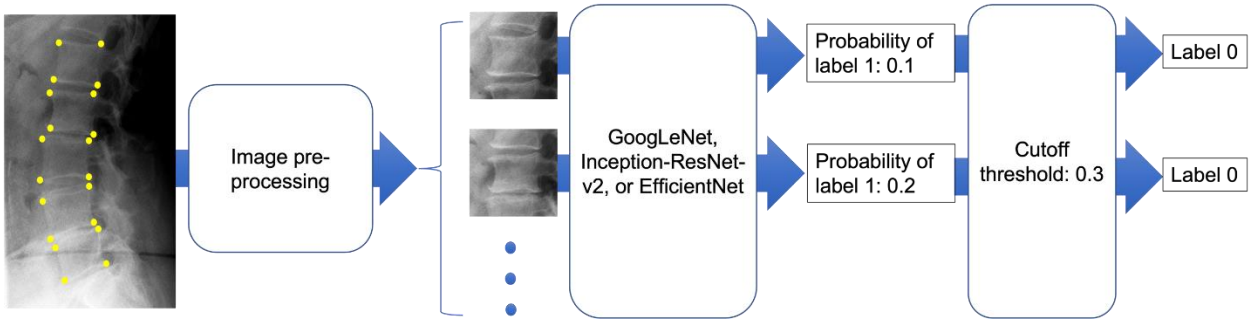
	Local dataset				MrOS dataset			
	Training set	Validation set	Test set	Total	Training set	Validation set	Test set	Total
m2ABQ	5,968	2,394	6,688	15,050	0	0	844	844
mSQ	NA				76,748	8,484	15,177	100,409

Model training

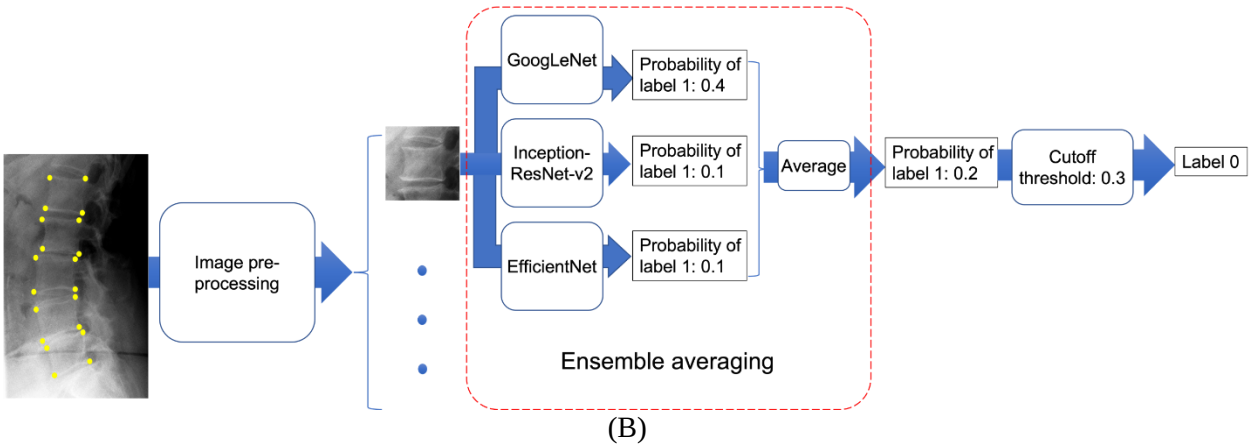
The inputs to each of our five models were the vertebral patches extracted from the spine radiographs by image pre-processing (described in Section A of the Supplemental Materials, which is similar to that in

(35)). The code for the image pre-processing is available at https://github.com/UW-CLEAR-Center/Preprocessing_for_Spinal_OCF_Detection_Multi_Datasets.

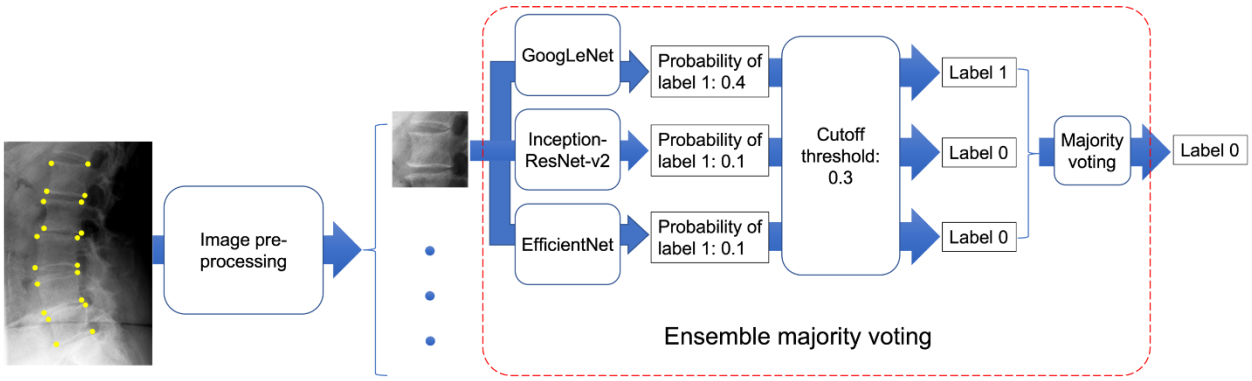
We trained five deep learning algorithms (see Figure 3), including GoogLeNet (44), Inception-ResNet-v2 (45), EfficientNet-B1 (46), and two ensemble algorithms. To train GoogLeNet, Inception-ResNet-v2, and EfficientNet-B1, transfer learning was used by pre-training a model on ImageNet (47) and fine-tuning the model on a target dataset. Besides this common transfer learning technique, we also built a model by first pre-training it on ImageNet, then tuning it on the MrOS-mSQ dataset, and finally fine-tuning it on the local-m2ABQ dataset. Recall that the local dataset contains clinical data, while the MrOS dataset was generated for research. To make the model performant on the clinical data, we finally fine-tuned each model on only the local-m2ABQ dataset rather than the combination of both the local-m2ABQ dataset and the MrOS-mSQ dataset. Since both the MrOS dataset and the local dataset contain vertebral patches, a model tuned on the MrOS-mSQ dataset before finally fine-tuned on the local-m2ABQ dataset can learn more relevant image features.



(A)



(B)



(C)

Figure 3. The flowchart of OCF classification using deep learning. Recall that the automatic image segmentation tool is a distinct body of work (see the INTRODUCTION section). In the current work, four manually annotated corner points of each vertebral body were used to extract the vertebral patch during image pre-processing. Taking an individual vertebral patch as an input, each of the five deep learning algorithms was used to build models to classify the vertebral patch to have label 0 or label 1. (A) shows the flowchart of OCF classification by GoogLeNet, Inception-ResNet-v2, or EfficientNet-B1. Each of these three models output a probability that the vertebral patch should be classified to have label 1. Then the vertebral patch was classified by comparing the probability and a pre-set cutoff threshold. (B) shows the flowchart of OCF classification by ensemble averaging, which averaged the probabilities output by the three individual models. Then the classification result was obtained by comparing the average probability and a pre-set cutoff threshold (details in the Model evaluation section of the MATERIALS AND METHODS section). (C) shows the flowchart of OCF classification by ensemble majority voting.

The classification result of ensemble majority voting was the majority classification result of the three individual models.

After training the models using the three individual algorithms mentioned above, two ensemble models were created using the ensemble averaging algorithm and the ensemble majority voting algorithm (see Figures 3(B) and 3(C)).

In summary, three deep learning models and two ensemble models were generated in each of the following three training tasks :

1) Task 1: Pre-train the model on ImageNet and fine-tune the model on the MrOS-mSQ dataset's training set (ImageNet → MrOS-mSQ).
2) Task 2: Pre-train the model on ImageNet and fine-tune the model on the local-m2ABQ dataset's training set (ImageNet → local-m2ABQ).
3) Task 3: The model tuned in Task 1 was further fine-tuned on the local-m2ABQ dataset's training set (ImageNet → MrOS-mSQ → local-m2ABQ).

In total, 15 models (5 models per task × 3 tasks) were built.

More details of model training are presented in Section B of the Supplemental Materials.

Model evaluation

Using both the local-m2ABQ dataset's test set and the MrOS-m2ABQ dataset's test set, we tested each of the 15 trained models described in the "Model training" section above. Each model trained in Task 1 was also tested on the MrOS-mSQ dataset's test set. All of the performance measures mentioned in this section were computed using the classification results on individual vertebral patches.

The ensemble majority voting algorithm does not output a numerical value on which a range of cutoff thresholds can be set (see Figure 3(C)). Thus, the AUC-PR and the AUC-ROC of the models built using the ensemble majority voting algorithm could not be computed. Instead, the following performance

measures were computed: accuracy, sensitivity, specificity, PPV, negative predictive value (NPV), false discovery rate (FDR=1–PPV), and F_1 score.

For the other trained models, all of the performance measures mentioned above were computed, including the AUC-PR and the AUC-ROC. For measures other than AUC-PR and AUC-ROC, a cutoff threshold was required. To set the cutoff threshold for each of these models, we used two thresholding methods, each applied to the validation set of the dataset whose training set was used to finally fine-tune the model. The same cutoff threshold was then used when testing the model on different test sets. The two thresholding methods are as follows:

1) Set the cutoff threshold to maximize the F_1 score. This automatically sets the cutoff threshold and balances the sensitivity and the PPV.
2) Manually set the threshold to make the PPV approximate 90%. Recall that we prioritize the PPV rather than the sensitivity for our opportunistic screening tool (see the INTRODUCTION section). Our initial consultation with local clinicians showed that a PPV of approximately 90% was appropriate.

The 95% confidence interval (CI) of each performance measure was computed using 2,000-fold bootstrap analysis.

IRB Approval

Retrieval of the local dataset was covered under the local retrospective institutional review board (IRB) for Diagnosis Radiology Images Deep Learning Project with a waiver of informed consent. For the MrOS dataset, at each medical center, a local IRB approved the MrOS study. All MrOS participants gave written informed consent at the time of the study.

RESULTS

Datasets

Table 2 shows the local dataset's metadata, including age, sex, race, ethnicity, radiograph generation year, and X-ray system vendor. In Section D of the Supplemental Materials, we also show the number and the percentage of the local dataset's radiographs generated by each type of machine. The MrOS dataset's metadata have been summarized in multiple publications (35-37, 42) and are shown in Table 3. Section E of the Supplemental Materials shows more details of the metadata of the MrOS dataset's training set.

Table 2. Metadata for the training, validation, and test sets of the local dataset, as well as the entire local dataset. The age data were retrieved from the RIS. The sex data were obtained from the DICOM metadata of the radiographs. The race and ethnicity data were retrieved from the electronic health record system. A subject could have multiple exams, which might not be from the same year. Consequently, multiple ages could be recorded for a subject. In each set, for every range of ages, we reported the number of recorded ages rather than the number of subjects. If a subject had multiple ages recorded, all of them were used to calculate the mean and the standard deviation.

	Training set	Validation set	Test set	Entire local dataset
	Number (percentage) of recorded ages			
Age at exam				
65-74	395 (53.2%)	181 (59.0%)	479 (56.6%)	1,055 (55.7%)
75-84	234 (31.6%)	84 (27.4%)	255 (30.1%)	573 (30.2%)
85-94	98 (13.2%)	32 (10.4%)	102 (12.0%)	232 (12.2%)
≥95	15 (2.0%)	10 (3.2%)	11 (1.3%)	36 (1.9%)
	Number			
Total recorded ages	742	307	847	1,896
	Mean ± standard deviation of ages in years			
Female	76 ± 9	75 ± 9	75 ± 8	75 ± 8
Male	75 ± 9	75 ± 9	75 ± 9	75 ± 9
All	75 ± 9	75 ± 9	75 ± 9	75 ± 9
	Number (percentage) of subjects			
Sex				

Female	339 (53.3%)	172 (56.0%)	467 (55.1%)	978 (54.6%)
Male	296 (46.5%)	135 (44.0%)	379 (44.8%)	810 (45.3%)
Not recorded	1 (0.2%)	0 (0%)	1 (0.1%)	2 (0.1%)
Race				
American Indian and Alaska Native	2 (0.3%)	2 (0.7%)	6 (0.7%)	10 (0.6%)
Asian	68 (10.7%)	37 (12.0%)	72 (8.5%)	177 (9.9%)
Black or African American	39 (6.2%)	20 (6.5%)	51 (6.0%)	110 (6.1%)
Native Hawaiian and Other Pacific Islander	2 (0.3%)	1 (0.3%)	3 (0.4%)	6 (0.3%)
White	474 (74.5%)	220 (71.7%)	654 (77.2%)	1348 (75.3%)
Multiple races	49 (7.7%)	25 (8.1%)	57 (6.7%)	131 (7.3%)
Not recorded	2 (0.3%)	2 (0.7%)	4 (0.5%)	8 (0.4%)
Ethnicity				
Hispanic or Latino	9 (1.4%)	5 (1.6%)	16 (1.9%)	30 (1.7%)
Not Hispanic or Latino	189 (29.7%)	138 (45.0%)	358 (42.3%)	685 (38.3%)
Not recorded	438 (68.9%)	164 (53.4%)	473 (55.8%)	1,075 (60.0%)
		Number		
Total subjects	636	307	847	1,790
		Number (percentage) of radiographs		
Radiograph generation year				
2000-2005	127 (17.1%)	49 (15.9%)	113 (13.3%)	289 (15.2%)
2006-2011	354 (47.7%)	135 (44.0%)	405 (47.8%)	894 (47.2%)
2012-2017	261 (35.2%)	123 (40.1%)	329 (38.9%)	713 (37.6%)
X-ray machine vendor				
Canon	5 (0.7%)	0 (0%)	5 (0.6%)	10 (0.5%)
DeJarnette Research Systems	48 (6.5%)	21 (6.8%)	48 (5.7%)	117 (6.2%)
Fujifilm	378 (50.9%)	157 (51.2%)	427 (50.4%)	962 (50.8%)
General Electric	202 (27.2%)	74 (24.1%)	232 (27.3%)	508 (26.8%)
Philips	104 (14.0%)	50 (16.3%)	127 (15.0%)	281 (14.8%)
Hybrid General Electric and Fujifilm	5 (0.7%)	5 (1.6%)	8 (1.0%)	18 (0.9%)
		Number		
Total radiographs	742	307	847	1,896

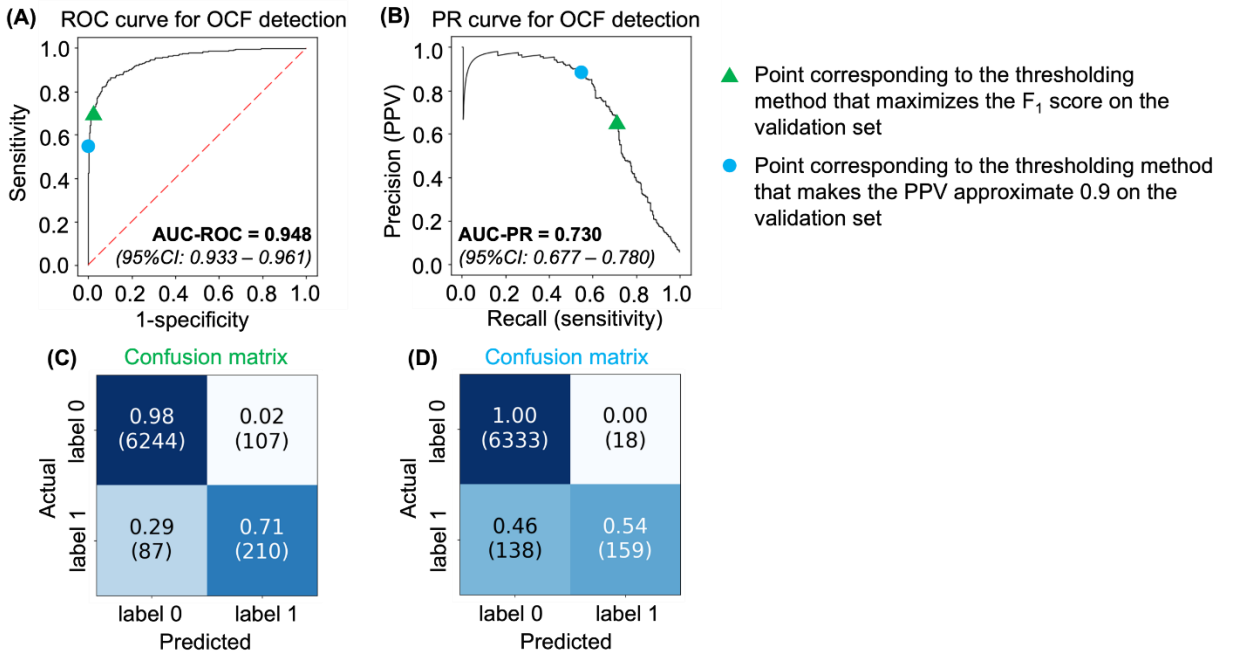
Table 3. Demographic information for the subjects in each of the entire, training, validation, and test sets from the MrOS dataset. By listing in the "Sampled test set (m2ABQ)" column, we also show the demographic information of the subjects in the sampled test set with 122 radiographs annotated by the m2ABQ criteria (see the "MrOS dataset" section of the "MATERIALS AND METHODS" section). The mean ± standard deviation of the ages was recorded at the baseline (Visit 1) and the follow-up (Visit 2) visits. If a subject reported multi-races, each race would be recorded.

	Training set	Validation set	Entire test set	Sampled test set (m2ABQ)	Entire dataset
	Mean ± standard deviation				
Age at Visit 1	73.7 ± 5.9	74.1 ± 6.2	73.5 ± 5.7	74.5 ± 5.8	73.7 ± 5.9
Age at Visit 2	77.8 ± 5.6	77.9 ± 5.6	77.5 ± 5.4	77.8 ± 5.3	77.7 ± 5.6
	Number (percentage) of subjects				
Race/ethnicity					
American Indian or Alaska Native	42 (0.8%)	7 (1.8%)	8 (1.2%)	0 (0.0%)	57 (0.9%)
Asian	159 (3.2%)	12 (3.1%)	25 (3.7%)	5 (4.8%)	196 (3.2%)
Black or African American	212 (4.2%)	21 (5.4%)	21 (3.1%)	4 (3.8)	254 (4.2%)
Hispanic or Latino	100 (2.0%)	11 (2.8%)	15 (2.2%)	0 (0.0%)	126 (2.1%)
Native Hawaiian or Other Pacific Islander	11 (0.2%)	3 (0.8%)	1 (0.1%)	1 (1.0%)	15 (0.2%)
White	4,492 (89.6%)	338 (86.1%)	611 (89.7%)	95 (90.4%)	5,441 (89.4%)

Model evaluation

We report the performance of our ensemble averaging algorithm in Tasks 2 (ImageNet → local-m2ABQ) and 3 (ImageNet → MrOS-mSQ → local-m2ABQ) in this manuscript and report the performance of the other models in Section C of the Supplemental Materials. The performance measures were computed using the classification results on individual vertebral patches.

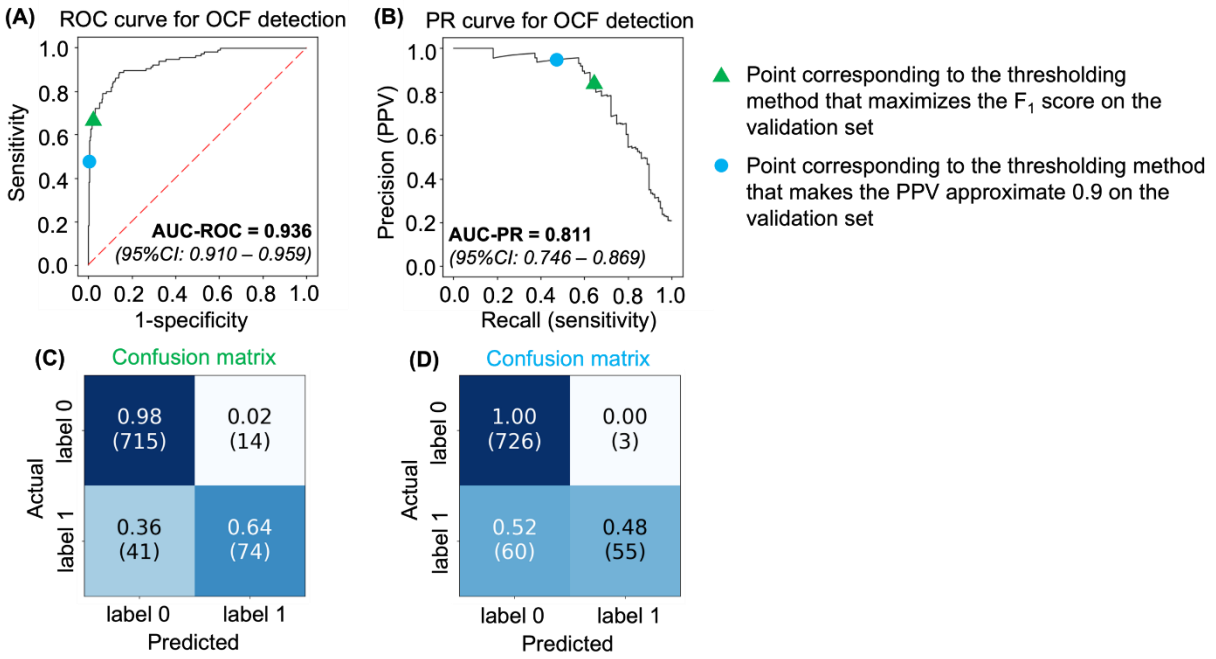
Figures 4 and 5 show the performance of the model built using the ensemble averaging algorithm in Task 2. Figures 4 and 5 show this model's performance on the local-m2ABQ dataset's test set and the MrOS-m2ABQ dataset's test set, respectively.



(E)

Cutoff threshold	Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
0.327	70.7% [65.3%-75.9%]	98.3% [98.0%-98.6%]	66.2% [60.9%-71.6%]	98.6% [98.3%-98.9%]	33.8% [28.4%-39.1%]	0.684 [0.630-0.737]	97.1% [96.7%-97.5%]
0.800	54.5% [48.1%-59.1%]	99.7% [99.6%-99.7%]	89.8% [85.1%-94.1%]	97.9% [97.5%-98.2%]	10.2% [5.9%-14.9%]	0.671 [61.5%-72.6%]	97.7% [97.3%-98.0%]

Figure 4. The performance of the model, which was built using the ensemble averaging algorithm in Task 2 and evaluated on the test set of the local-m2ABQ dataset. (A) The ROC curve and the AUC-ROC with its 95% CI. (B) The PR curve and the AUC-PR with its 95% CI. (C) When the cutoff threshold (0.327) is set to maximize the F_1 score on the local-m2ABQ dataset's validation set, the confusion matrix with the number of vertebral bodies in each of the four cells shown in the parentheses. (D) The confusion matrix when the cutoff threshold (0.800) is manually set to make the PPV approximate 90% on the local-m2ABQ dataset's validation set. (E) Using each thresholding method, the sensitivity, specificity, PPV, NPV, FDR, F_1 score, and accuracy with their 95% CIs.



(E)

Cutoff threshold	Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
0.327	64.3% [55.3%-72.8%]	98.1% [97.1%-99.0%]	84.1% [75.9%-91.5%]	94.6% [92.9%-96.2%]	15.9% [8.5%-24.1%]	0.729 [0.640-0.811]	93.5% [91.8%-95.1%]
0.800	47.8% [39.6%-56.8%]	99.6% [99.0%-100%]	94.8% [88.9%-100%]	92.4% [90.5%-94.1%]	5.2% [0-11.1%]	0.636 [54.7%-72.4%]	92.5% [90.9%-94.3%]

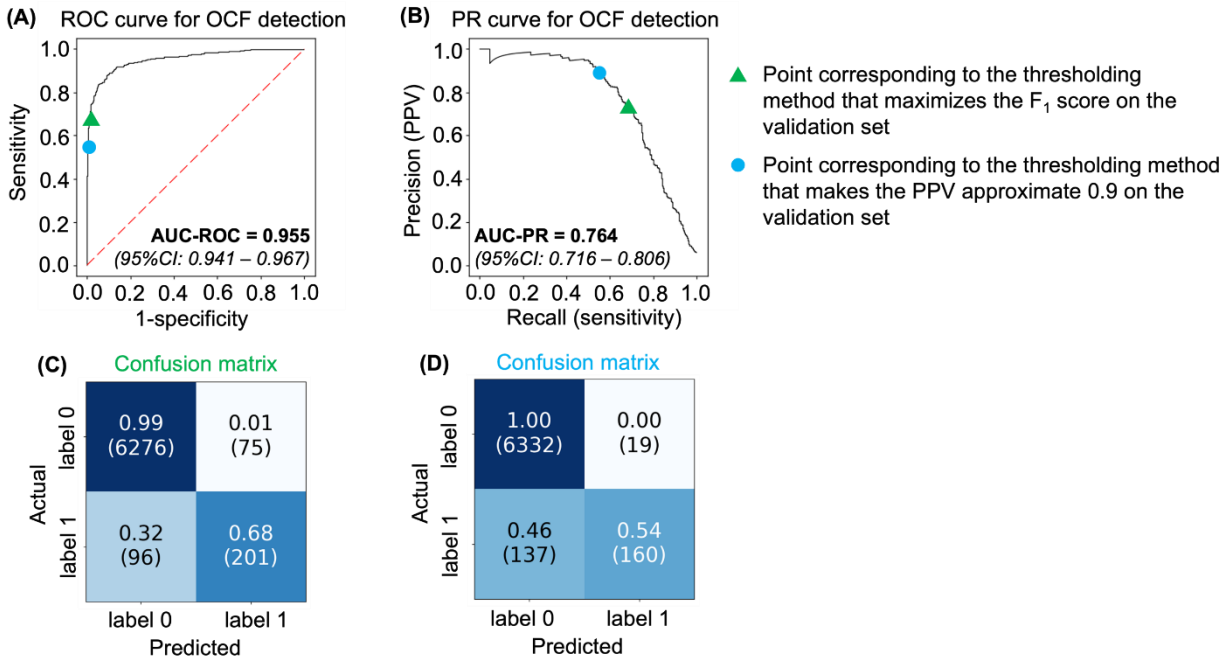
Figure 5. The performance of the model, which was built using the ensemble averaging algorithm in Task 2 and evaluated on the test set of the MrOS-m2ABQ dataset. (A) The ROC curve and the AUC-ROC with its 95% CI. (B) The PR curve and the AUC-PR with its 95% CI. (C) When the cutoff threshold (0.327) is set to maximize the F_1 score on the local-m2ABQ dataset's validation set, the confusion matrix with the number of vertebral bodies in each of the four cells shown in the parentheses. (D) The confusion matrix when the cutoff threshold (0.800) is manually set to make the PPV approximate 90% on the local-m2ABQ dataset's validation set. (E) Using each thresholding method, the sensitivity, specificity, PPV, NPV, FDR, F_1 score, and accuracy with their 95% CIs.

On the local-m2ABQ dataset's test set, the model mentioned above yielded an AUC-ROC of 0.948 and an AUC-PR of 0.730. After setting the cutoff threshold to make the PPV approximate 90% on the local-

m2ABQ dataset's validation set, this model achieved a sensitivity of 54.5%, a specificity of 99.7%, a PPV of 89.8%, an NPV of 97.9%, an FDR of 10.2%, an F_1 score of 0.671, and an accuracy of 97.7%.

On the MrOS-m2ABQ dataset's test set, the model mentioned above yielded an AUC-ROC of 0.936 and an AUC-PR of 0.811. After setting the cutoff threshold to make the PPV approximate 90% on the local-m2ABQ dataset's validation set, this model achieved a sensitivity of 47.8%, a specificity of 99.6%, a PPV of 94.8%, an NPV of 92.4%, an FDR of 5.2%, an F_1 score of 0.636, and an accuracy of 92.5%.

Figure 6 shows the performance of the model built using the ensemble averaging algorithm in Task 3 and evaluated on the local-m2ABQ dataset's test set. This model yielded an AUC-ROC of 0.955 and an AUC-PR of 0.764. After setting the cutoff threshold to make the PPV approximate 90% on the local-m2ABQ dataset's validation set, this model achieved a sensitivity of 53.9%, a specificity of 99.7%, a PPV of 89.4%, an NPV of 97.9%, an FDR of 10.6%, an F_1 score of 0.672, and an accuracy of 97.7%.



(E)

Cutoff threshold	Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
0.764	67.7% [62.2%-73.0%]	98.8% [98.6%-99.1%]	72.8% [67.7%-78.1%]	98.5% [98.2%-98.8%]	27.2% [21.9%-32.3%]	0.702 [0.648-0.755]	97.4% [97.1%-97.8%]
0.900	53.9% [48.4%-59.5%]	99.7% [99.6%-99.8%]	89.4% [84.7%-93.8%]	97.9% [97.5%-98.2%]	10.6% [6.2%-15.3%]	67.2% [61.6%-72.3%]	97.7% [97.2%-98.0%]

Figure 6. The performance of the model, which was built using the ensemble averaging algorithm in Task 3 and evaluated on the test set of the local-m2ABQ dataset. (A) The ROC curve and the AUC-ROC with its 95% CI. (B) The PR curve and the AUC-PR with its 95% CI. (C) When the cutoff threshold (0.764) is set to maximize the F_1 score on the local-m2ABQ dataset's validation set, the confusion matrix with the number of vertebral bodies in each of the four cells shown in the parentheses. (D) The confusion matrix when the cutoff threshold (0.900) is manually set to make the PPV approximate 90% on the local-m2ABQ dataset's validation set. (E) Using each thresholding method, the sensitivity, specificity, PPV, NPV, FDR, F_1 score, and accuracy with their 95% CIs.

Comparison between the models

For each deep learning algorithm, there were three training tasks. Each model was tested using two or three test sets. Table 4 shows the F_1 score, the AUC-PR, and the AUC-ROC of each (deep learning algorithm, training task, test set) combination. In this section, each model's cutoff threshold was set to maximize the F_1 score on the corresponding validation set.

Table 4. F_1 scores, AUC-PR, and AUC-ROC for each (deep learning algorithm, training task, test set) combination. The AUC-PR and the AUC-ROC of the models built using the ensemble majority voting algorithm could not be computed (see the "Model evaluation" section of the "MATERIALS AND METHODS" section). In this table, yellow and magenta are used to mark the MrOS dataset and the local dataset, respectively.

Training task	**Task 1**: ImageNet → MrOS-mSQ			**Task 2**: ImageNet → local-m2ABQ		**Task 3**: ImageNet → MrOS-mSQ → local-m2ABQ	
Test set	MrOS-mSQ	MrOS-m2ABQ	Local-m2ABQ	MrOS-m2ABQ	Local-m2ABQ	MrOS-m2ABQ	Local-m2ABQ
				F_1 score			
GoogLeNet	0.751	0.691	0.579	0.698	0.668	0.694	0.701
Inception-ResNet-V2	0.729	0.652	0.523	0.670	0.659	0.698	0.674

EfficientNet-B1	0.743	0.667	0.543	0.705	0.650	0.747	0.689
Ensemble averaging	0.773	**0.677**	**0.566**	**0.729**	0.684	**0.761**	0.702
Ensemble majority voting	**0.776**	0.648	0.553	0.706	**0.694**	0.713	**0.712**

AUC-PR

GoogLeNet	0.817	0.782	0.606	0.784	0.698	0.804	0.736
Inception-ResNet-V2	0.798	0.795	0.636	0.809	0.656	0.801	0.696
EfficientNet-B1	0.816	**0.796**	0.628	0.785	0.703	0.808	0.746
Ensemble averaging	**0.841**	**0.796**	**0.658**	**0.811**	**0.730**	**0.831**	**0.764**

AUC-ROC

GoogLeNet	0.990	0.897	0.918	0.927	0.941	0.933	0.949
Inception-ResNet-V2	**0.993**	**0.925**	0.914	0.930	0.925	0.922	0.947
EfficientNet-B1	**0.993**	0.914	0.916	0.914	0.941	0.933	**0.958**
Ensemble averaging	0.992	0.911	**0.930**	**0.936**	**0.948**	**0.940**	0.955

In the local dataset's test set and the MrOS dataset's test set, the percentages of vertebral bodies with label 1 are 4.5% (computed using the table at the bottom of Figure 2) and 1.1% (35), respectively. Because AUC-ROC is less suitable than AUC-PR for a highly imbalanced test set (48), the models are compared not using AUC-ROC but using the F_1 score and the AUC-PR.

DISCUSSION

The number of subjects in each of the training and test sets was determined by striking the balance between obtaining a large set and reducing manual annotation time. A large set is more likely to contain diverse data. Thus, a large training set can reduce model overfitting. A large test set can ensure accurate measures of model performance. However, since manual annotation is time-consuming, we could not wait to train and test our models after annotating a very large number of radiographs.

The ensemble averaging model trained in Task 2 achieved our pre-specified objectives of AUC-PR >0.70 and AUC-ROC >0.90 on both the local dataset and the MrOS dataset. When setting the cutoff threshold to make the PPV approximately 90% on the local-m2ABQ dataset's validation set, we obtained high PPVs and specificities with moderate sensitivities on both datasets. This is acceptable for our clinical use case of an opportunistic screening tool described in the INTRODUCTION section, in which the PPV and specificity rather than the sensitivity should be prioritized. An opportunistic screening tool could be clinically useful with a moderate sensitivity and a high specificity or PPV. Given the volume of radiographic exams that cover some portion of the thoracic and lumbar spine at most medical institutions, it is prudent to consider the downstream effects of positive and negative predictive results. A positive predictive result would result in provider efforts guiding the patient to the appropriate clinical care as well as patient expense, worry, radiation exposure, and potential harm. A negative predictive result would result in no further action and would not affect the current standard of clinical care. Our opportunistic screening tool will only augment current clinical practice rather than replace radiologist interpretation or any other step in the current clinical workflow. In this setting, a false negative is a missed opportunity, but could still be possibly caught by the current standard of care. A false positive triggers extra work that has no obvious benefit to the patient but potential harm and financial burden. Our model with a PPV of about 90% and a sensitivity of about 50% can detect nearly half of the unreported fractured vertebral bodies with limited extra cost. It is worth noting that many diagnostic tests in use today have modest sensitivities. Papanicolaou smear has a sensitivity of 55.4% and a specificity of 94.6% (49).

In the "Comparison between the models" section of the RESULTS section, we compared the performance of each (deep learning algorithm, training task, test set) combination. We have six observations:

1) In each (training task, test set) combination, the models built using the two ensemble algorithms typically outperformed the other models.
2) In each (training task, test set) combination, the two ensemble algorithms typically produced models with similar F_1 scores. Unlike the ensemble majority voting algorithm that outputs categorical values,

the ensemble averaging algorithm provided numerical outputs to which different cutoff thresholds could be applied. Thus, the ensemble averaging algorithm is more flexible and can be adapted for different clinical use cases.

3) In Task 2 (ImageNet → local-m2ABQ), the model built using the ensemble averaging algorithm had a better F_1 score and a higher AUC-PR on the MrOS-m2ABQ dataset than on the local-m2ABQ dataset. This shows that the model built using the ensemble averaging algorithm has some generalizability. Counterintuitively, this model performed worse on the test set of the local-m2ABQ dataset, whose training set was used for fine-tuning this model, than on the MrOS-m2ABQ dataset. The reason could be that the data in the local dataset are more diverse, especially in subject positioning and image artifacts, increasing difficulty of OCF classification.
4) On each test set, each model trained in Task 3 (ImageNet → MrOS-mSQ → local-m2ABQ) typically had a higher F_1 score and a better AUC-PR than the corresponding model trained in Task 2 (ImageNet → local-m2ABQ) did. Our transfer learning technique in Task 3 could improve models' performance. However, since each model trained in Task 3 was tuned using both datasets, we cannot claim that this model is generalizable. We need more datasets to show these models' generalizability.
5) In Task 1 (ImageNet → MrOS-mSQ), the AUC-PR of each model tested on the MrOS-mSQ dataset was higher than that of each model tested on the MrOS-m2ABQ dataset but to a limited degree (e.g., 5.7% for the ensemble averaging algorithm). This could imply that our two binary OCF labeling systems (simplified from the mSQ criteria and the m2ABQ criteria, respectively) are similar.
6) In Task 1, the F_1 score and the AUC-PR of each model tested on the MrOS-mSQ dataset were higher than those of each model tested on the local-m2ABQ dataset, respectively (e.g., 36.6% and 27.8% greater, respectively by the F1 score and the AUC-PR, for the ensemble averaging algorithm). The models fine-tuned on the MrOS-mSQ dataset were not generalizable to the local-m2ABQ dataset. The MrOS dataset was obtained for research, while the local dataset was extracted from clinical data

that were more diverse in demographics, X-ray techniques, and image artifact variations. This greater diversity is likely the cause of poor performance by models only fine-tuned on the MrOS dataset.

Researchers from other research projects (25-29) reported approaches to automatically detecting OCFs using radiographs. Using lumbar or thoracolumbar spine radiographs, Chou et al. (25) did automatic segmentation to extract the vertebral bodies and classified each vertebral body using an ensemble method. Using similar methods, Li et al. (26) trained models to automatically detect vertebral fractures on lateral spine radiographs. Chen et al. (27) and Murata et al. (28) respectively trained a deep learning model to detect vertebral fractures on a radiograph without vertebral body segmentation. The main limitation of each of the above projects is that a single-site dataset was used. This resulted in a more homogeneous population, making the trained models less generalizable.

Xiao et al. (29) trained and tested their models on women's lateral spine and chest radiographs from multiple sites, showing that their models had good generalizability and could serve as an opportunistic screening tool for female OCF screening. Based on their models, they developed a software program with a user interface. However, except for two datasets, they did not mention the source, the dataset construction process, and the demographic information of the other datasets in detail. The two known datasets were retrieved from the Osteoporotic fractures in women (MsOS) Hong Kong dataset (50). Like the MrOS dataset, the MsOS Hong Kong dataset was originally collected for research and has some selection bias. Their recruitment criteria included that all subjects were able to walk without assistance (50). The radiographs in this dataset likely contain far fewer imaging chain artifacts like angulation, position, overlapping, motion, and equipment, which are commonly seen in standard clinical imaging, and are seen when comparing the local and MrOS datasets in our study.

In contrast to the above projects, we used data assembled from multiple sites with detailed description of the dataset construction process and demographic information (see Figure 2 and Table 2 describing the local dataset, as well as the papers (35-37, 42) describing the MrOS dataset). Our local dataset was retrieved from local clinical sites and thus is more consistent with the distribution of clinical data. Shown in Table 2, the local dataset contains subjects that have varied race, ethnicity, and gender, as well as

radiographs generated from different X-ray machines, which could help improve the generalizability of our trained models.

Our models have several limitations:

1) We used lateral spine radiographs to build our classifiers. This type of radiograph is optimized to show bones, and thus a rational initial target for research. However, to increase the target population in the future, other radiographs like lateral chest or abdominal radiographs should be used.
2) Our current model classifies individual vertebral bodies extracted from spine radiographs using manual annotation. This ensures that the vertebral bodies are correctly bounded on a radiograph but is not automated or scalable. As mentioned in the INTRODUCTION section, we are testing and separately reporting image segmentation models to automatically localize the vertebral bodies on a radiograph.
3) Currently, we only have one dataset (the local dataset) containing data acquired in varied clinical settings for diagnostic purposes. The number of annotated radiographs in the local dataset is small. We need more annotated clinical data to train our model and test its generalizability. In the future, we will annotate more radiographs from various clinical sites using semi-automated approaches.
4) In this study, the cutoff thresholds set using the two thresholding methods might not be the best for the clinical use case. We have already surveyed a variety of clinical providers to determine an acceptable performance threshold for automated opportunistic OCF screening. We will further analyze our survey results to determine the most appropriate cutoff threshold for the clinical use case.
5) In this study, we did not analyze incorrectly classified cases and explore how image features contribute to each model's outputs. These two tasks should be implemented in the future to understand how the model works, its failure modes, and how to further improve the model.

In conclusion, we used five deep learning algorithms to train models that detected OCFs of vertebral bodies extracted from spine radiographs. The ensemble averaging model trained in Task 2 achieved our pre-specified objectives of AUC-PR >0.70 and AUC-ROC >0.90 on both the local dataset and the MrOS

dataset. This model has good performance and some generalizability and can serve as a critical component of our future automated opportunistic screening tool.

Reference

1. Looker AC, Borrud LG, Dawson-Hughes B, et al. Osteoporosis or low bone mass at the femur neck or lumbar spine in older adults: United States, 2005-2008. *NCHS Data Brief* 2012;93.
2. Kanis JA, on behalf of the World Health Organization Scientific Group (2007). Assessment of osteoporosis at the primary health-care level. Technical Report. WHO Collaborating Centre for Metabolic Bone Diseases, University of Sheffield, UK, 2007.
3. Hodsman AB, Leslie WD, Tsang JF, et al. 10-year probability of recurrent fractures following wrist and other osteoporotic fractures in a large clinical cohort: an analysis from the Manitoba Bone Density Program. *Arch Intern Med* 2008;168(20):2261-2267.
4. Roux S, Cabana F, Carrier N, et al. The World Health Organization Fracture Risk Assessment Tool (FRAX) underestimates incident and recurrent fractures in consecutive patients with fragility fractures. *J Clin Endocrinol Metab* 2014;99(7):2400-2408.
5. Robinson CM, Royds M, Abraham A, et al. Refractures in patients at least forty-five years old: a prospective analysis of twenty-two thousand and sixty patients. *J Bone Joint Surg Am* 2002;84(9):1528-1533.
6. Center JR, Nguyen TV, Schneider D, et al. Mortality after all major types of osteoporotic fracture in men and women: an observational study. *Lancet* 1999;353(9156):878-882.
7. Meadows ES, Whangbo A, McQuarrie N, et al. Compliance with mammography and bone mineral density screening in women at least 50 years old. *Menopause* 2011;18(7):794-801.
8. King AB, Fiorentino DM. Medicare payment cuts for osteoporosis testing reduced use despite tests' benefit in reducing fractures. *Health Aff* 2011;30(12):2362-2370.

9. Jain S, Bilori B, Gupta A, et al. Are men at high risk for osteoporosis underscreened? A quality improvement project. *Perm J* 2016;20(1):60-64.
10. Pickhardt PJ, Pooler BD, Lauder T, et al. Opportunistic screening for osteoporosis using abdominal computed tomography scans obtained for other indications. *Ann Intern Med* 2013;158(8):588-595.
11. Anderson PA, Polly DW, Binkley NC, et al. Clinical use of opportunistic computed tomography screening for osteoporosis. *J Bone Joint Surg* 2018;100(23):2073-2081.
12. Alacreu E, Moratal D, Arana E. Opportunistic screening for osteoporosis by routine CT in Southern Europe. *Osteoporos Int* 2017;28(3):983-990.
13. Li YL, Wong KH, Law MW, et al. Opportunistic screening for osteoporosis in abdominal computed tomography for Chinese population. *Arch Osteoporos* 2018;13(1):1-7.
14. Cheng X, Zhao K, Zha X, et al. Opportunistic screening using low‐dose CT and the prevalence of osteoporosis in China: a nationwide, multicenter study. *J Bone Miner Res* 2021:36(3):427-435.
15. Fang Y, Li W, Chen X, et al. Opportunistic osteoporosis screening in multi-detector CT images using deep convolutional neural networks. *Eur Radiol* 2021;31(4):1831-1842.
16. Nam KH, Seo I, Kim DH, et al. Machine learning model to predict osteoporotic spine with hounsfield units on lumbar computed tomography. *J Korean Neurosurg Soc* 2019;62(4):442-449.
17. Löffler MT, Jacob A, Scharr A, et al. Automatic opportunistic osteoporosis screening in routine CT: improved prediction of patients with prevalent vertebral fractures compared to DXA. *Eur Radiol* 2021;31:6069-6077.
18. Yasaka K, Akai H, Kunimatsu A, et al. Prediction of bone mineral density from computed tomography: application of deep learning with a convolutional neural network. *Eur Radiol* 2020;30:3549-3557.
19. Bar A, Wolf L, Amitai OB, et al. Compression fractures detection on CT. In: Proceedings of SPIE Medical Imaging: Computer-Aided Diagnosis, Orlando, FL. International Society for Optics and Photonics, 2017; 1013440.

20. Yilmaz EB, Buerger C, Fricke T, et al. Automated Deep Learning-Based Detection of Osteoporotic Fractures in CT Images. In: Proceedings of Machine Learning in Medical Imaging, Strasbourg, France. Cham, Switzerland: Springer, 2021; 376-385.
21. Husseini M, Sekuboyina A, Bayat A, et al. Conditioned variational auto-encoder for detecting osteoporotic vertebral fractures. In: Proceedings of the International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging, Granada, Spain. Cham, Switzerland: Springer, 2019; 29-38.
22. Tomita N, Cheung YY, Hassanpour S. Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. *Comput Biol Med* 2018;98:8-15.
23. Lee S, Choe EK, Kang HY, et al. The exploration of feature extraction and machine learning for predicting bone density from simple spine X-ray images in a Korean population. *Skeletal Radiol* 2020;49(4):613-618.
24. Zhang B, Yu K, Ning Z, et al. Deep learning of lumbar spine X-ray for osteopenia and osteoporosis screening: A multicenter retrospective cohort study. *Bone* 2020;140:115561.
25. Chou PH, Jou THT, Wu HTH, et al. Ground truth generalizability affects performance of the artificial intelligence model in automated vertebral fracture detection on plain lateral radiographs of the spine. *Spine J* 2022;22(4):511-523.
26. Li YC, Chen HH, Horng-Shing Lu H, et al. Can a deep-learning model for the automated detection of vertebral fractures approach the performance level of human subspecialists? *Clin Orthop Relat Res* 2021;479(7):1598-1612.
27. Chen HY, Hsu BW, Yin YK, et al. Application of deep learning algorithm to detect and visualize vertebral fractures on plain frontal radiographs. *PLoS One* 2021;16(1):e0245992.
28. Murata K, Endo K, Aihara T, et al. Artificial intelligence for the detection of vertebral fractures on plain spinal radiography. *Sci Rep* 2020;10(1):20031.

29. Xiao BH, Zhu MSY, Du EZ, et al. A software program for automated compressive vertebral fracture detection on elderly women's lateral chest radiograph: Ofeye 1.0. *Quant Imaging Med Surg* 2022;12(8):4259-4271.
30. IMV reports general X-ray procedures growing at 5.5% per year, as number of installed X-ray units declines. CISION PRWeb.
31. Bolotin HH. DXA in vivo BMD methodology: an erroneous and misleading research and clinical gauge of bone mineral status, bone fragility, and bone remodelling. *Bone* 2007;41(1):138-154.
32. Kim TY, Schafer AL. Variability in DXA reporting and other challenges in osteoporosis evaluation. *JAMA Intern Med* 2016;176(3):393-395.
33. Carberry GA, Pooler BD, Binkley N, et al. Unreported vertebral body compression fractures at abdominal multidetector CT. *Radiology* 2013;268(1):120-126. https://www.prweb.com/releases/2011/2/prweb8127064.htm. Accessed April 12, 2022.
34. Renslo J, Chang B, Dong Q, et al. U-Net for spine segmentation – towards osteoporotic fracture detection. Accepted by the ASNR meeting 2023.
35. Dong Q, Luo G, Lane NE, et al. Deep learning classification of spinal osteoporotic compression fractures on radiographs using an adaptation of the Genant semiquantitative criteria. *Acad Radiol* 2022;29(12):1819-1832.
36. Orwoll E, Blank JB, Barrett-Connor E, et al. Design and baseline characteristics of the osteoporotic fractures in men (MrOS) study-a large observational study of the determinants of fracture in older men. *Contemp Clin Trials* 2005;26(5):569-585.
37. Blank JB, Cawthon PM, Carrion-Petersen ML, et al. Overview of recruitment for the osteoporotic fractures in men study (MrOS). *Contemp Clin Trials* 2005;26(5):557-568.
38. Dong Q, Luo G, Haynor D, et al. DicomAnnotator: a configurable open-source software program for efficient DICOM image annotation. *J Digit Imaging* 2020;33(6):1514-1526.
39. Aaltonen HL, O'Reilly MK, Linnau K, et al. m2ABQ – a proposed refinement of the modified algorithm-based qualitative classification of osteoporotic vertebral fractures. *Osteoporos Int* 2022.

40. Lentle BC, Berger C, Probyn L, et al. Comparative analysis of the radiology of osteoporotic vertebral fractures in women and men: cross-sectional and longitudinal observations from the Canadian multicentre wsteoporosis study (CaMos). *J Bone Miner Res* 2018;33(4):569-579.
41. LireNLPSystem package documentation. GitHub. https://github.com/UW-CLEAR-Center/LireNLPSystem. Accessed November 9, 2022.
42. Cawthon PM, Haslam J, Fullman R, et al. Methods and reliability of radiographic vertebral fracture detection in older men: the osteoporotic fractures in men study. *Bone* 2014;67:152-155.
43. Genant HK, Wu CY, van Kuijk C, et al. Vertebral fracture assessment using a semiquantitative technique. *J Bone Miner Res* 1993;8(9):1137-1148.
44. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: Proceedings of CVPR, Boston, MA. Washington, D.C.: IEEE Computer Society, 2015; 1-9.
45. Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In: Proceedings of AAAI, San Francisco, CA. Palo Alto, CA: AAAI Press, 2017; 4278-4284.
46. Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of ICML, Long Beach, CA. JMLR.org: 2019; 6105-6114.
47. Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In: Proceedings of CVPR, Miami, FL. Washington, D.C.: IEEE Computer Society, 2009; 248-255.
48. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Proceedings of ICML, Pittsburgh, PA. New York, NY: Association for Computing Machinery, 2006; 233-240.
49. Kripke C. Pap smear vs. HPV screening tests for cervical cancer. *Am Fam Physician* 2008;77(12):1740-1742.
50. Wáng YXJ, Deng M, Griffith JF, et al. 'Healthier Chinese spine': an update of osteoporotic fractures in men (MrOS) and in women (MsOS) Hong Kong spine radiograph studies. *Quant Imaging Med Surg* 2022;12(3):2090-2105.

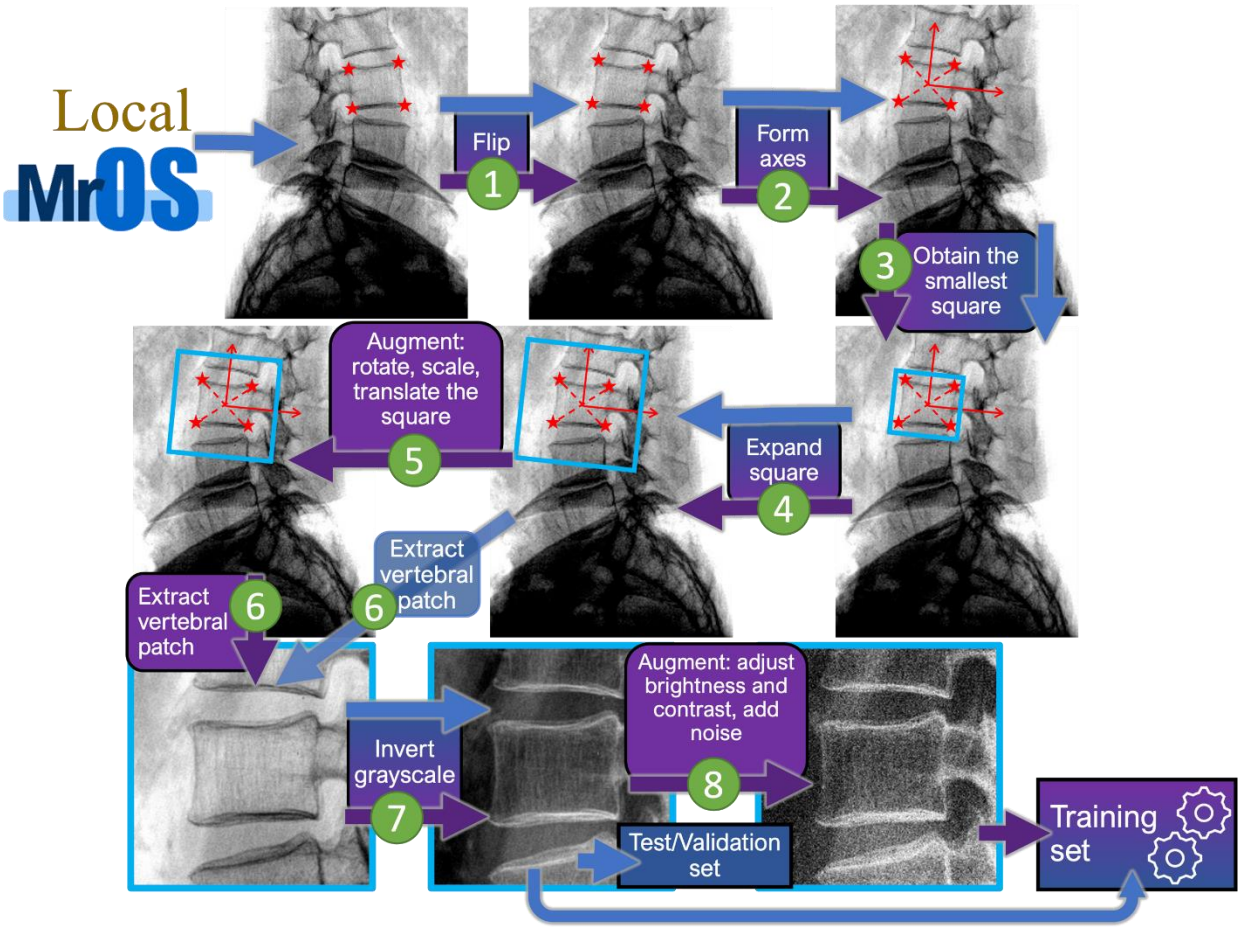
Supplemental Materials

A. Image pre-processing and augmentation

A.1 Overview of image pre-processing and augmentation

Before feeding the data instances into a model, the image preprocessing and augmentation steps were conducted, as shown in Supplementary Figure 1. Both the radiographs in the MrOS dataset and the radiographs in the local dataset were processed using all of these steps. The purposes of image preprocessing and augmentation are summarized as follows:

1) Extract the vertebral patches from the spine radiographs.
2) Ensure that the heterogeneity among vertebral patches is within a moderate range. Excessive heterogeneity among the data can make the classification task more complex. However, too little heterogeneity can make the trained model have poor generalizability.



Supplementary Figure 1. Steps of image pre-processing and augmentation. These steps were applied to each vertebral body in a radiograph. To extract each vertebral body, the four corner points represented by the red stars were used. The purple and blue arrows demonstrate vertebral patch extraction with and without image augmentation, respectively. Only the vertebral patches in the training set were augmented. The training set included both the original vertebral patches and the augmented ones. We did not augment the vertebral patches in the test and validation sets. The steps for extracting a vertebral patch are as follows: 1) to conform to the convention, convert the radiograph to a 16-bit image and flip the radiograph horizontally, if needed, to make the subject face left; 2) create the coordinate system with the x-axis bisecting the angle between the two diagonals connecting the four corner points; 3) draw the smallest square bounding box, which can cover all of the four corner points and whose edges are parallel to the coordinate axes; 4) expand the square bounding box from its center to enlarge the area fourfold, to avoid

cutoff of part of the vertebral body while adding surrounding image context; 5) scale, rotate, and translate the square to augment the vertebral patch; 6) extract the vertebral patch; 7) invert the grayscale of the vertebral patch if the bones are darker than the background; 8) augment the vertebral patch by adjusting the brightness/contrast as well as adding Gaussian noise; and 9) resize the vertebral patch to 224×224 pixels for training GoogLeNet or 299×299 pixels for training Inception-ResNet-v2 and EfficientNet-B1.

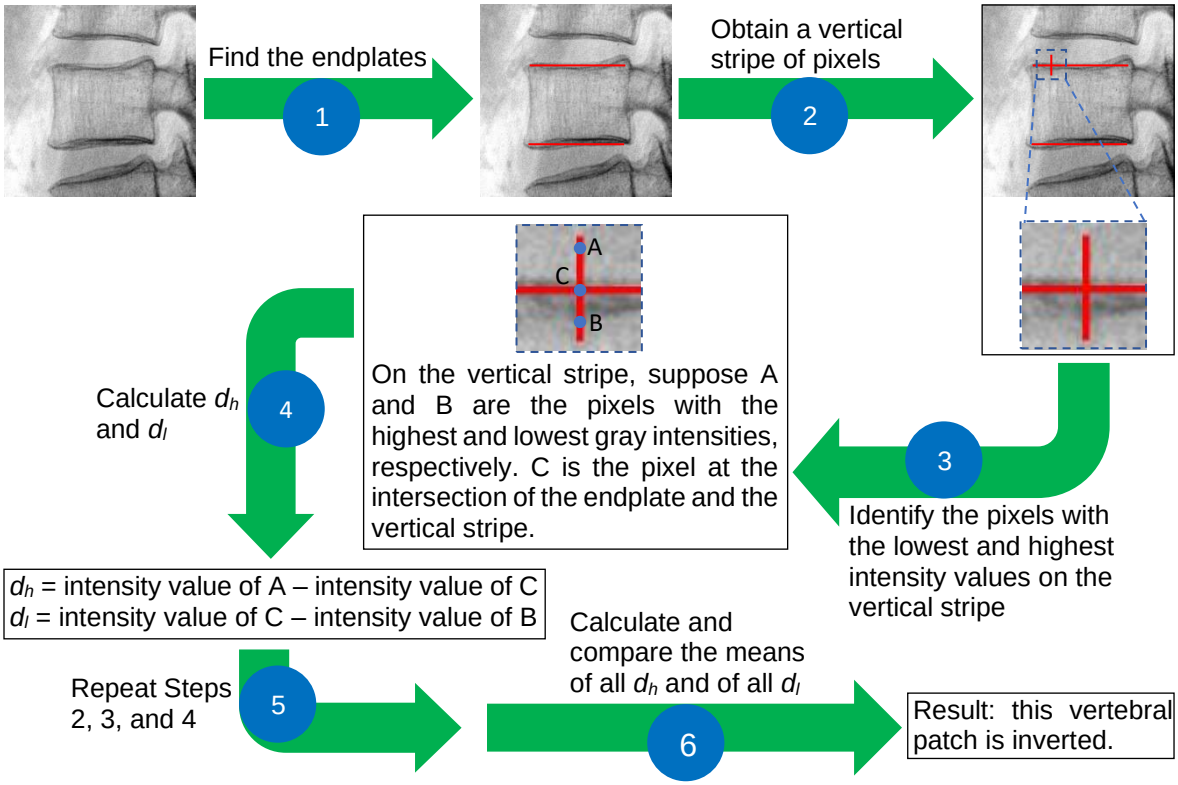
Image augmentation steps were applied only to the training set, whereas the image pre-processing steps were applied to all of the training, validation, and test sets. To balance the training set of the local dataset, one augmented patch was generated for each vertebral patch labeled with label 0, while eight augmented patches were generated for each vertebral patch labeled with label 1. Because the local dataset was small, we did not balance the training set of the local dataset by downsampling the data instances in the majority class (label 0). By contrast, the MrOS dataset is relatively large; therefore, the data instances in the majority class were downsampled to balance the training set of the MrOS dataset. Since the MrOS dataset's training set was balanced, five augmented patches were generated for each original vertebral patch in the training set of the MrOS dataset.

The details of the image pre-processing and augmentation steps were the same as those in our previous work (35), except for Step 7, in which the grayscale of the vertebral patch was inverted if the bones were darker than the background (see Supplementary Figure 1). In Section A.2, the new algorithm developed for Step 7 is introduced. In Section A.3, the hyper-parameter values of this new algorithm are determined. In Section A.4, the testing results of the new algorithm is presented.

A.2 Determining whether bones are darker than the background in a vertebral patch

An inverted patch is a vertebral patch in which bones are darker than the background. An algorithm for detecting inverted patches was presented our previous work (35). However, the algorithm was built and tested using only the MrOS dataset. The algorithm in our previous work (35) did not work well on the vertebral patches of the local dataset, as the vertebral patches tended to be noisier and less clear than those

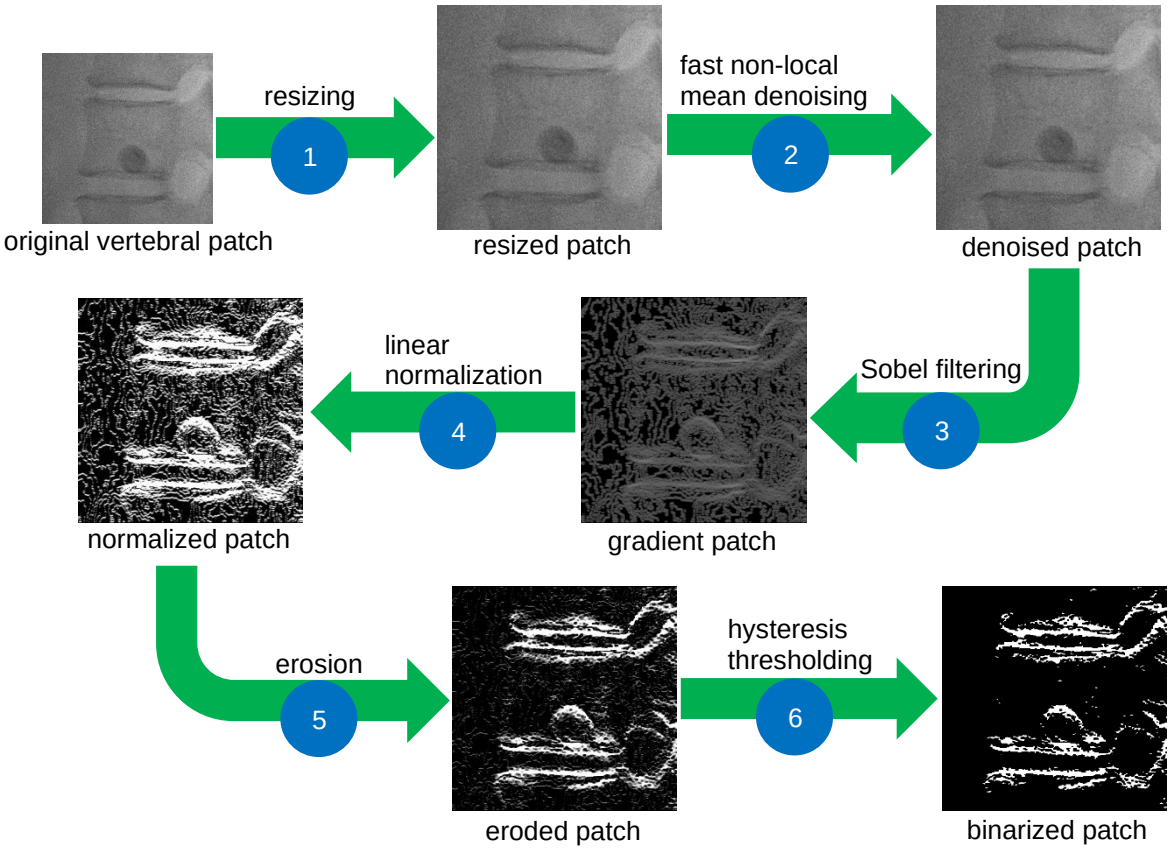
in the MrOS dataset. The new algorithm introduced in this section is a modification of the algorithm in our previous work (35) based on the following intuition. If the vertebral patch is inverted, then the endplates tend to be dark, and thus their gray intensities are closer to the lower end of the image histogram. Otherwise, if the vertebral patch is not inverted, then the endplates tend to be bright, and their gray intensities are closer to the high end of the image histogram. Supplementary Figure 2 outlines the general steps for detecting inverted patches. In our approach, the details in Steps 2–6 are the same as those in our previous work (35), whereas Step 1 is modified. In Step 1, the endplates of the vertebral body were located. Note that, in a vertebral patch, the vertebral body is not tilted much (see Supplementary Figure 1); thus, the endplates are approximately horizontal. In our previous work (35), Sobel filtering (51) and hysteresis thresholding (51) were used to find the horizontal edges representing the endplates of the vertebral patches in the MrOS dataset. As the vertebral patches in the local dataset are noisier than those in the MrOS dataset, only applying Sobel filtering and hysteresis thresholding to the vertebral patches of the local dataset can lead to many horizontal edges because of the noise rather than the endplates. The horizontal edges resulting from noise, in subsequent steps, can preclude the identification of inverted vertebral patches. Thus, reducing noise and horizontal edges resulting from noise is critical.



Supplementary Figure 2. Steps for determining whether a vertebral patch is inverted: 1) find the endplates; 2) on the endplate, draw a vertical strip of pixels such that the midpoint is the pixel on one of the endplates found in the first step; 3) on the vertical stripe, identify the pixels with the lowest and highest intensity values; 4) calculate d_l and d_h; 5) traverse the pixels on the endplates found in the first step and repeat Steps 2-4 to obtain all d_l and all d_h; and 6) calculate the mean of all d_l and the mean of all d_h, and then compare the two means to decide whether the vertebral patch is inverted. If the mean of all d_l is < the mean of all d_h, the gray intensities of the endplates are closer to the low end of the image histogram. The endplates tend to be dark and thus the vertebral patch is inverted. Otherwise, if the mean of all d_l is ≥ the mean of all d_h, the vertebral patch is not inverted.

Supplementary Figure 3 illustrates our modified technique for identifying the endplates in each vertebral patch. First, all vertebral patches were resized to the same size. For each resized vertebral patch, fast non-local means denoising (52, 53) was used to reduce noise. Sobel filtering (51) was applied to

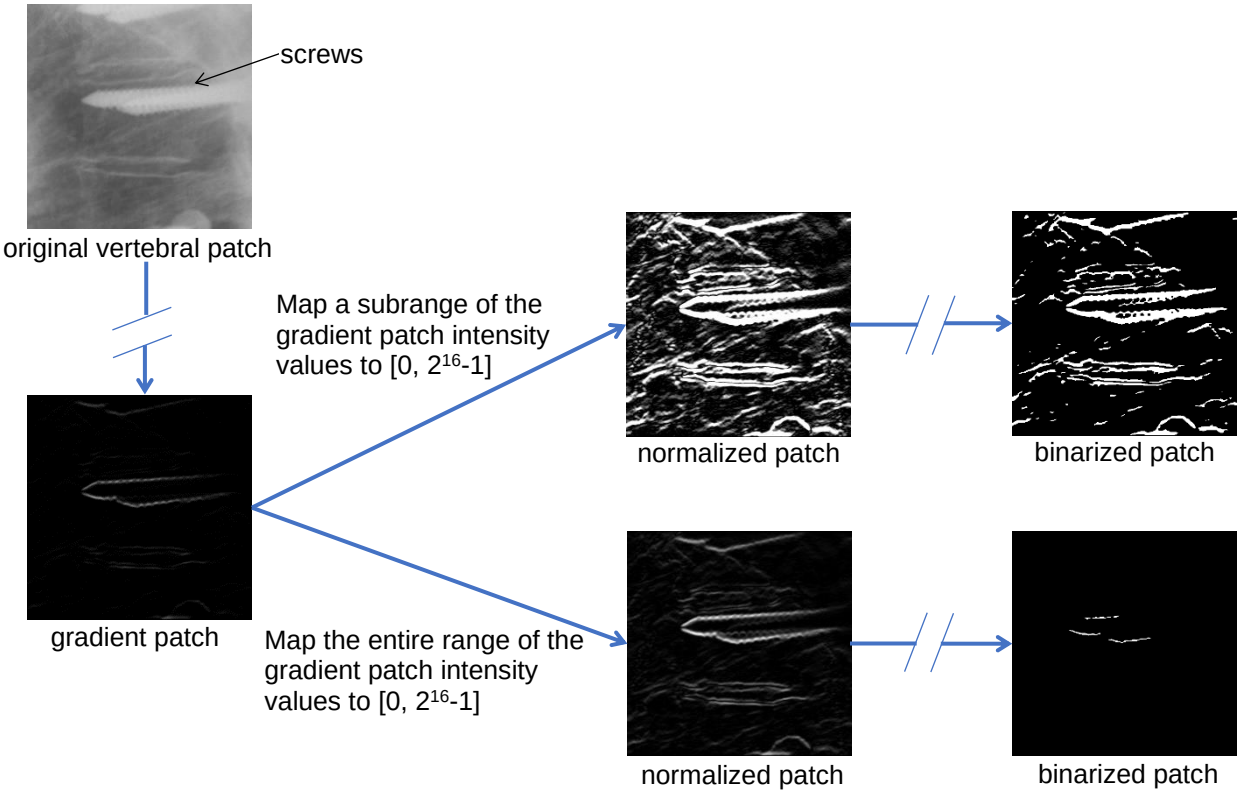
calculate the gradient of each pixel in the denoised vertebral patch. The patch formed from all pixel gradients is referred to as a gradient patch (see Supplementary Figure 3). The absolute values of these gradients were calculated to turn the negative intensity values in the gradient patch to positive. The pixels had a higher intensity value on the detected horizontal edges than in other areas of the gradient patch. Linear normalization (54) was conducted on the gradient patch to increase the contrast of the gradient patch. On the normalized patch, horizontal edges resulting from noise were eroded by applying the erosion operation (51, 55), using a square kernel. Finally, hysteresis thresholding (51) was applied to the eroded patch. The hyper-parameters involved in these steps are introduced in Section A.3.



Supplementary Figure 3. Steps to find the endplates of a vertebral body: 1) resizing the original vertebral patch; 2) fast non-local means denoising to reduce noise; 3) Sobel filtering to obtain the gradient of the

denoised patch; 4) linear normalization to adjust the contrast of the gradient patch; 5) erosion on the normalized patch; and 6) hysteresis thresholding to binarize the eroded patch and obtain the endplates.

Some details of the linear normalization (Step 4 in Supplementary Figure 3) must be mentioned. The aim in conducting linear normalization is to increase the contrast between the endplates and background on the gradient patch. With this contrast increasing, the endplates can be more easily detected using the subsequent steps. Typically, linear normalization maps the entire range of the image intensity values to a new range (54). In our case, a subrange of the intensity values in each gradient patch was linearly mapped to the new range $[0, 2^{16}-1]$, which is the widest range for 16-bit images. The intensity values above and below this subrange were mapped to $2^{16}-1$ or 0, respectively. The endpoints of this subrange were hyper-parameters that required tuning, which are introduced in Section A.3. The reason for mapping a subrange rather than the entire range of the gradient patch intensity values was to ensure that the contrast between the endplates and background would increase to a much greater extent. In some gradient patches, the brightest area might not include the endplates (see the gradient patch in Supplementary Figure 4, where the brightest areas are the edges of the screws). In this case, if the entire range of the intensity values of the gradient patch is mapped to the new range $[0, 2^{16}-1]$, the contrast between the endplates and the background could increase only slightly, and it may not be possible to detect the endplates (see Supplementary Figure 4).



Supplementary Figure 4. Comparison between the two strategies of linear normalization. The original vertebral patch contains screws, which are used to treat various conditions, including scoliosis. If the entire range of the gradient patch intensity values is mapped to [0, 2^{16}-1] (the lower path of the figure), then only a part of the screws' edges remains in the binarized patch. In the upper path of the figure, a subrange of the gradient patch intensity values is mapped to [0, 2^{16}-1]. The pixel intensity values above and below this subrange are mapped to 2^{16}-1 and 0, respectively. Thus, the endplates can also remain in the binarized patch.

After finding the endplates, Steps 2-6 in Supplementary Figure 2 were used to check whether the vertebral patch is inverted. After determining whether each vertebral patch in the radiograph was inverted, majority voting was used to determine whether the radiograph was inverted. In case of a tie in the majority voting, the radiograph was randomly assigned to "inverted" or "non-inverted." After majority

voting, each vertebral patch in this radiograph was regarded as having the same inverting status as the radiograph, regardless of the inverting status label of the vertebral patch before majority voting.

A.3 Hyper-parameters in the entire process of inverted radiograph detection

In the entire process of inverted radiograph detection (see Supplementary Figures 2 and 3), 12 hyper-parameter values are determined. The hyper-parameter that determines the number of iterations of the erosion operation on each normalized patch was assigned the default value of 1 (55). Supplementary Table 1 lists the other 11 hyper-parameters, which were tuned to determine the optimal values. To tune these hyper-parameters, 92 radiographs were randomly selected from the training set of the local dataset. Then, a 2,000-round random search was conducted. In each round, for each of the 11 hyper-parameters, a value was randomly sampled from within the search range, as listed in Supplementary Table 1. With the hyper-parameter values sampled, the entire process of inverted radiograph detection was applied to each of the 92 radiographs. Among all 2,000 search rounds, the hyper-parameter values resulting in the best accuracy constituted our final result, as listed in Supplementary Table 1.

Supplementary Table 1. The 11 hyper-parameters for determining whether a radiograph is inverted. We list the algorithm step in which each hyper-parameter is involved, the hyper-parameter definition, the random search range, and the optimal value.

Step	**Hyper-parameter**	**Definition**	**Search range**	**Optimal value**
resizing the vertebral patch	**size**	Size in pixels of the resized vertebral patch.	{224, 225, …, 1024}	505
fast non-local mean denoising	**h**	Hyper-parameter determining the filtering strength.	{7, 8, …, 13}	8
	template window size	Size in pixels of the template window.	{3, 5, …, 15}	5
	search window size	Size in pixels of the window into which the template window is slid to calculate the weighted average.	{11, 13, …, 31}	25
Sobel filtering	**Sobel kernel size**	Size in pixels of the Sobel kernel.	{3, 5, …, 31}	7

linear normalization	**left endpoint**	Left endpoint of the subrange of the gradient patch intensity values that are mapped to [0, 2^{16}-1].	[0, 0.3]	0.29
	right endpoint	Right endpoint of the subrange of the gradient patch intensity values that are mapped to [0, 2^{16}-1].	[0.7, 1]	0.89
erosion	**erosion kernel size**	Size in pixels of the square kernel for conducting the erosion operation.	{1, 3, …15}	3
hysteresis thresholding	**low threshold**	Threshold used to determine which pixels should be turned to black.	[0.4, 0.8]	0.59
	high threshold	Threshold used to determine which pixels should be turned to white.	(low threshold, 1]	0.77
putting vertical stripes on each pixel of the detected horizontal lines	**length**	Length in pixels of the vertical stripe.	{11, 13, …, 101}	75

A.4 Testing the algorithm for inverted radiograph detection

First, the proposed algorithm for inverted radiograph detection was tested on the 92 radiographs that were used for hyper-parameter tuning. The resulting accuracy was 98.9%. Supplementary Table 2(A) lists the corresponding confusion matrices. Second, to test the generalizability of our algorithm, another 94 radiographs were randomly selected from the local dataset, and our algorithm was applied to each of them. The accuracy of the algorithm on these 94 radiographs was 91.5%. Supplementary Table 2(B) presents the resulting confusion matrix.

Supplementary Table 2. Confusion matrices after applying the proposed algorithm for inverted radiograph detection to (A) 92 radiographs used for hyper-parameter tuning and (B) 94 radiographs used for testing.

		Predicted	
		Non-inverted	**Inverted**
Actual	**Non-inverted**	51	0
	Inverted	1	40

(A)

		Predicted	
		Non-inverted	**Inverted**
Actual	**Non-inverted**	46	1
	Inverted	7	40

(B)

B. Details of model training

To build models for OCF classification, we built the five deep learning algorithms (see Figure 3 in the main body of the paper), namely, GoogLeNet, Inception-ResNet-v2, EfficientNet-B1, and the two ensemble algorithms. These algorithms were implemented using Python 3.7.6, TensorFlow 2.4.1 (56), and TF-Slim 1.1.0 (57).

We chose each of these three individual algorithms because of the combination of the following two factors:

1) The performance of the algorithm in published benchmark analysis.
2) The number of parameters that the algorithm has. Considering that we have very limited data instances, too many parameters could cause overfitting.

In each of the three training tasks (see the "Model training" section of the "MATERIALS AND METHODS" section in the main body of the paper), we only needed to train GoogLeNet, Inception-ResNet-v2, and EfficientNet-B1. Given the classification results of these three models after training, the classification result of each ensemble model was computed (see Figures 3(B) and 3(C) in the main body of the paper). Thus, in all three training tasks, we built 15 deep learning models, 9 of which through model training. In the remainder of this section, the details of the model training are provided.

To boost the performance of the models, we applied transfer learning (58), which consists of the following steps: 1) pre-training a model on a large source dataset and 2) fine-tuning the weights of the model on the target dataset. ImageNet (47) is one of our source datasets. The TensorFlow Model Garden (59) provides GoogLeNet and Inception-ResNet-v2 models pre-trained on ImageNet. The online open-source code (60) provides an EfficientNet-B1 model pretrained on ImageNet. Before fine-tuning a model that was pre-trained on ImageNet, the output layer of the corresponding neural network was adjusted for binary classification to fit our OCF classification. This modified output layer could not be initialized using the weights of the model pretrained on ImageNet. Instead, this output layer was initialized using He

initialization (61). When tuning each model that was pre-trained on ImageNet, we conducted the following two steps:

1) Freeze all layers except the output layer. Train the model with a fixed learning rate of 10^{-3}, a batch size of 20, Adam optimization (62), and 15 epochs.
2) Keep freezing several layers close to the input layer and unfreeze the other layers. Use a smaller learning rate to continue to tune the model. The batch size was 20.

Some details of the second step are shown in the following:

1) The number of frozen layers was a hyper-parameter.
2) Learning rate decay was applied to the weights of the unfrozen layers. If the AUC-PR of the model evaluated after each epoch on the validation set did not increase in any of the subsequent two epochs, the learning rate was multiplied by a decay factor of <1. Once the learning rate was decayed, the neural network with the weights leading to the best validation result ever obtained during training this model was reloaded. Subsequently, the decayed learning rate was used to continue the model training process. Both the learning rate when model training starts (termed the initial learning rate) and the decay factor were hyper-parameters.
3) The weighted cross-entropy loss function (63) was used to penalize false positives and false negatives in different ways. The factor controlling the false-negative weight was set to 1. The factor controlling the false-positive weight (denoted by pos_weight) was a hyper-parameter.
4) To avoid overfitting, early stopping (64) was used during model training. If the model AUC-PR evaluated after each epoch on the validation set did not increase in any of the subsequent 10 epochs, model training was ended.

A three-step transfer learning technique was also designed (see Task 3 in the "Model training" section of the "MATERIALS AND METHODS" section in the main body of the paper). Before finally fine-tuning a model on the local-m2ABQ dataset, the model was already tuned on the training set of the MrOS-mSQ dataset. Recall that before tuning the model on the training set of the MrOS-mSQ dataset, the mSQ categories were simplified into two classes. Thus, the model tuned on the training set of the MrOS-

mSQ dataset was already used for binary classification. All model weights tuned on the training set of the MrOS-mSQ dataset were used to initialize the model in the fine-tuning step. When fine-tuning each model that was already tuned on the MrOS-mSQ dataset, we froze several layers close to the input layer and used a small learning rate to fine-tune the model. The batch size was 20. The learning rate decay, the weighted cross-entropy loss function, and the early stopping were used. The number of frozen layers *L*, the initial learning rate, the decay rate, and pos_weight were hyper-parameters that required tuning.

In each step of each model tuning process, dropout (64) was used to avoid overfitting. For each unit in the fully connected layer before the output layer, the probability of dropping this unit is a hyper-parameter referred to as the dropout rate.

In summary, in the training process of each model, five hyper-parameters required tuning. These hyper-parameters, listed in Supplementary Table 3, were tuned by random search (64) for 2,000 rounds, with the goal of maximizing the AUC-PR on the validation set. The initial learning rate, decay factor, and pos_weight were determined on the logarithmic scale. The dropout rate was determined on a linear scale. The deepest frozen layer *L* was searched on a list. The search space of *L* was different for GoogLeNet, Inception-ResNet-v2, and EfficientNet-B1 because they have distinct architectures. The code representing each layer of each neural network is provided in its open-source code (60, 65, 66) and original paper (44-46). Supplementary Table 4 lists the optimal values of these hyper-parameters for each model training process. The hyper-parameters not mentioned in this section were set to their default values given by the original papers (44-46) and open-source code (60, 65, 66) of these deep learning algorithms.

Supplementary Table 3. Five hyper-parameters that were tuned by random search.

Hyper-parameter	**Description**	**Search range or search space**
Initial learning rate	Learning rate when model training begins.	$[10^{-6}, 10^{-3}]$
Decay factor	Value by which the learning rate is multiplied to decrease the learning rate.	$[10^{-3}, 1]$
pos_weight	Factor controlling the false-positive weight in the weighted cross entropy loss function.	[0, 10]

Dropout rate	Probability of dropping each unit of the fully connected layer before the output layer.	[0, 1]
L	Deepest frozen layer. If *L* ≠ none, the input layer up to *L* are frozen. Otherwise, if *L* = none, no layer is frozen.	GoogLeNet: {none, 1a, 2b, 2c} Inception-ResNet-v2: {none, 1a, 2a, 2b, 3b, 4a} EfficientNet-B1: {none, stem, block1, block2, block3, block4, block5}

Supplementary Table 4. For each GoogLeNet, Inception-ResNet-v2, and EfficientNet-B1, the optimal value of each hyper-parameter in each training task.

		Optimal value		
		Task 1: ImageNet → MrOS-mSQ	Task 2: ImageNet → local-m2ABQ	Task 3: ImageNet → MrOS-mSQ → local-m2ABQ
GoogLeNet	Initial learning rate	6.95×10^{-4}	5.43×10^{-4}	3.02×10^{-4}
	Decay factor	8.53	16.03	654.70
	pos_weight	0.14	0.35	0.29
	Dropout rate	0.25	0.48	0.33
	L	None	1a	1a
Inception-ResNet-v2	Initial learning rate	2.2×10^{-4}	2.90×10^{-4}	1.53×10^{-4}
	Decay factor	71.65	5.17	1.87
	pos_weight	0.71	6.48	0.64
	Dropout rate	0.70	0.42	0.24
	L	1a	1a	None
EfficientNet-B1	Initial learning rate	7.96×10^{-3}	2.85×10^{-3}	1.65×10^{-4}
	Decay factor	14.02	1.71	14.06
	pos_weight	0.14	0.31	1.69
	Dropout rate	0.94	0.20	0.97
	L	block1	block3	stem

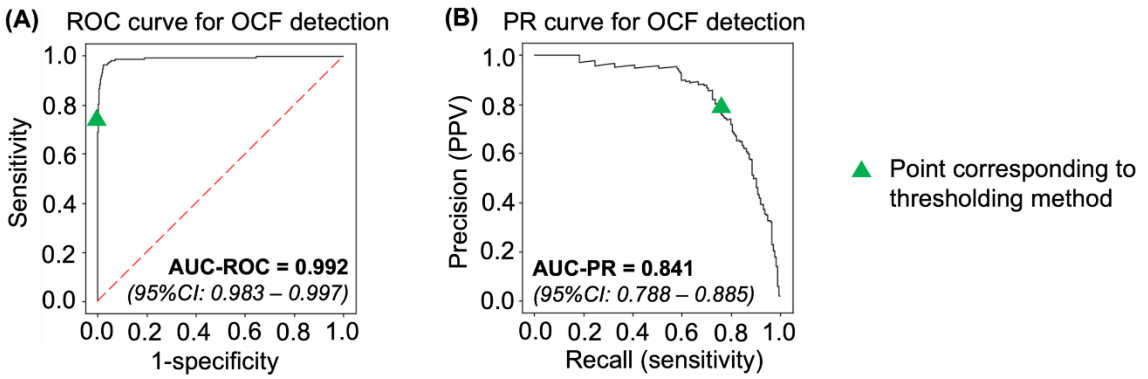
Hyper-parameter tuning was performed on two Ubuntu Linux servers concurrently:1) Xeon E5-2630 with four Nvidia GeForce TITAN Xp GPUs and 512 GB of memory and 2) Xeon Gold 5215 with four Nvidia GeForce 2080 Ti GPUs and 96 GB of memory. Each of the models (including those built using the two ensemble algorithms) was evaluated on the first server using one GPU.

C. Performance of all trained models

Each of Supplementary Figures 5-39 shows the performance of a (deep learning algorithm, training task, test set) combination. In each figure, the following performance measures are included whenever possible:

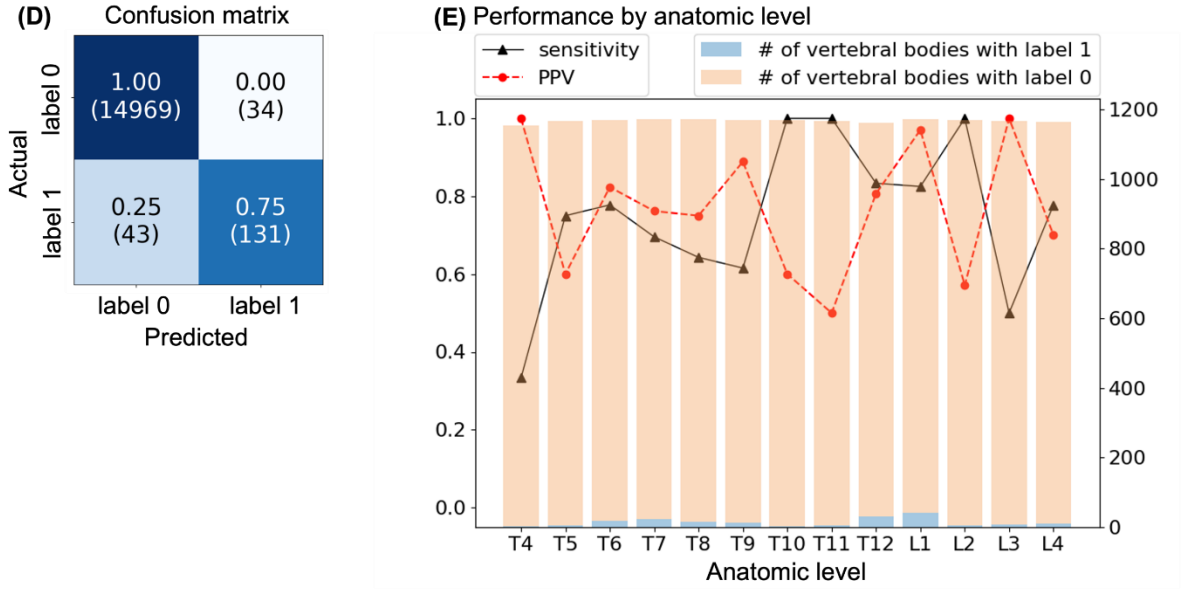
1) The ROC curve and AUC-ROC with its 95% CI.
2) The PR curve and AUC-PR with its 95% CI.
3) A table listing sensitivity, specificity, PPV, NPV, FDR, F_1 score, and accuracy with 95% CIs when setting the cutoff threshold to maximize the F_1 score on the validation set.
4) The confusion matrix using the same cutoff threshold as above.
5) A subfigure showing the PPV and sensitivity at each anatomic level of the spine. The PPVs at some levels could not be computed, as the denominator of the PPV at each of these levels (i.e., the number of vertebral bodies predicted as label 1 at each level) was 0.

Recall that when the model is built using the ensemble majority voting algorithm, the ROC curve and the PR curve cannot be drawn.

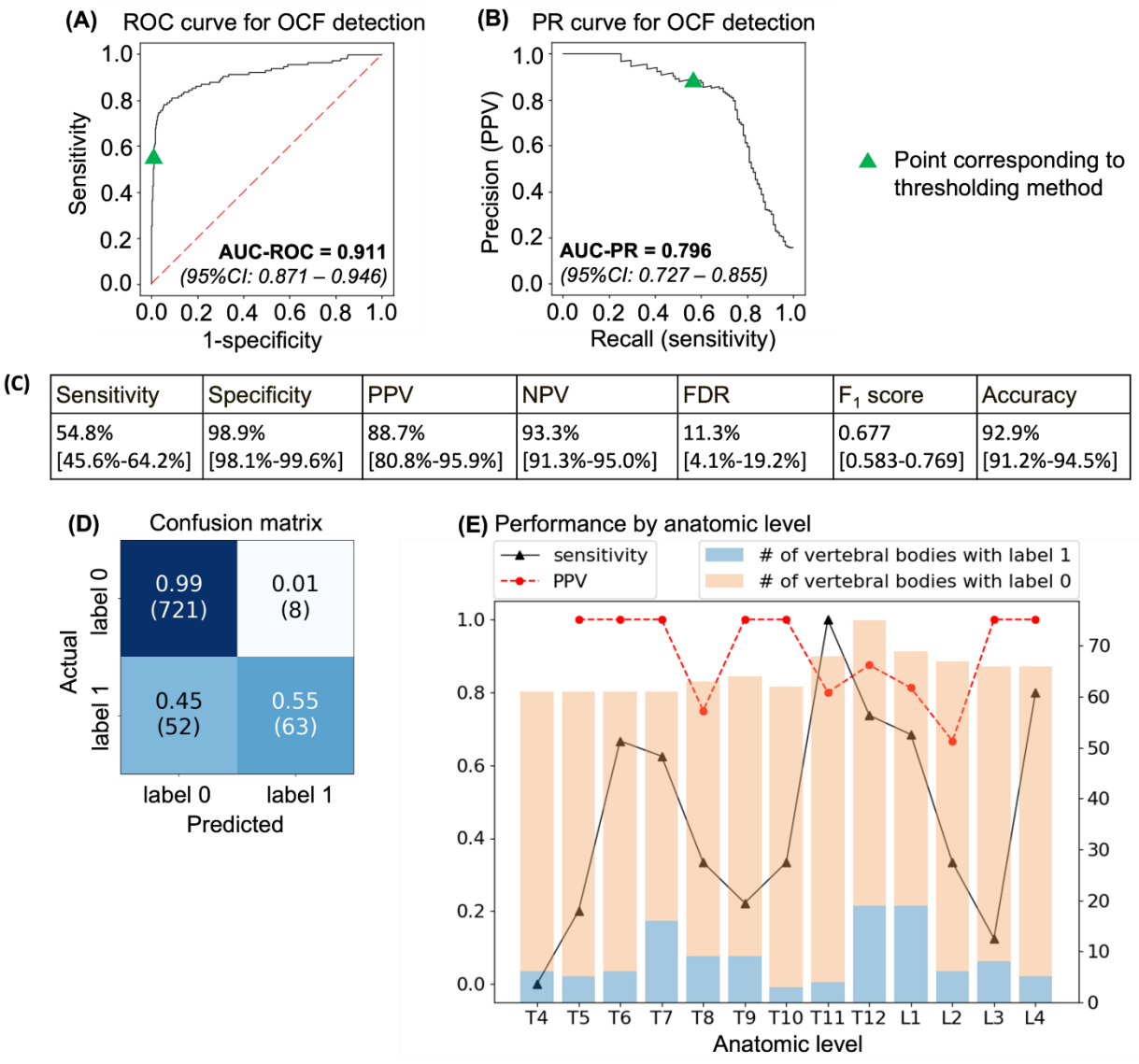


(C)

Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
75.3% [68.8%-81.3%]	99.8% [99.7%-99.8%]	79.4% [73.1%-85.3%]	99.7% [99.6%-99.8%]	20.6% [14.7%-26.9%]	0.773 [0.709-0.833]	99.5% [99.4%-99.6%]

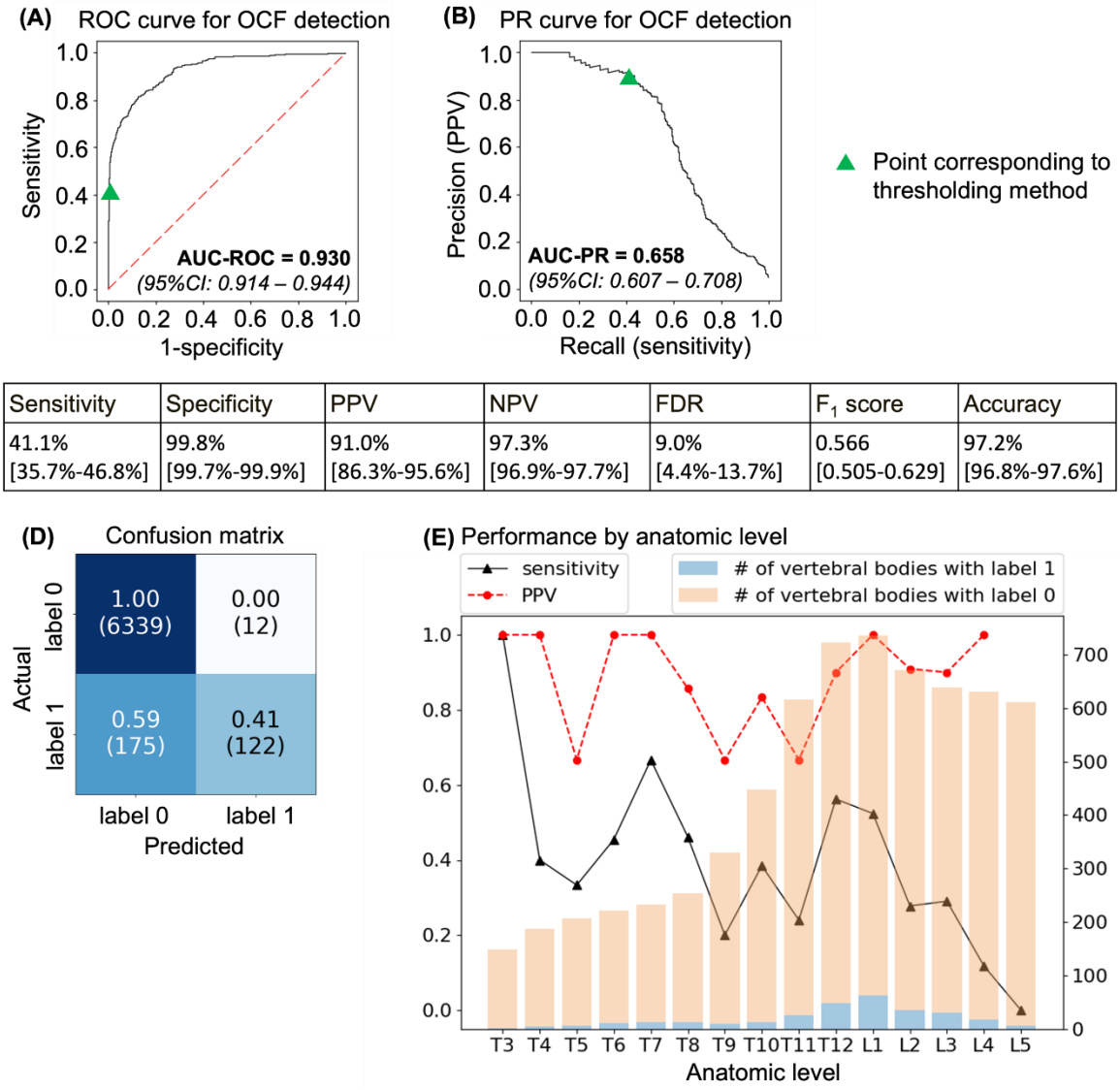


Supplementary Figure 5. Performance of the model built using the ensemble averaging algorithm in Task 1 and evaluated on the test set of the MrOS-mSQ dataset.



Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
54.8% [45.6%-64.2%]	98.9% [98.1%-99.6%]	88.7% [80.8%-95.9%]	93.3% [91.3%-95.0%]	11.3% [4.1%-19.2%]	0.677 [0.583-0.769]	92.9% [91.2%-94.5%]

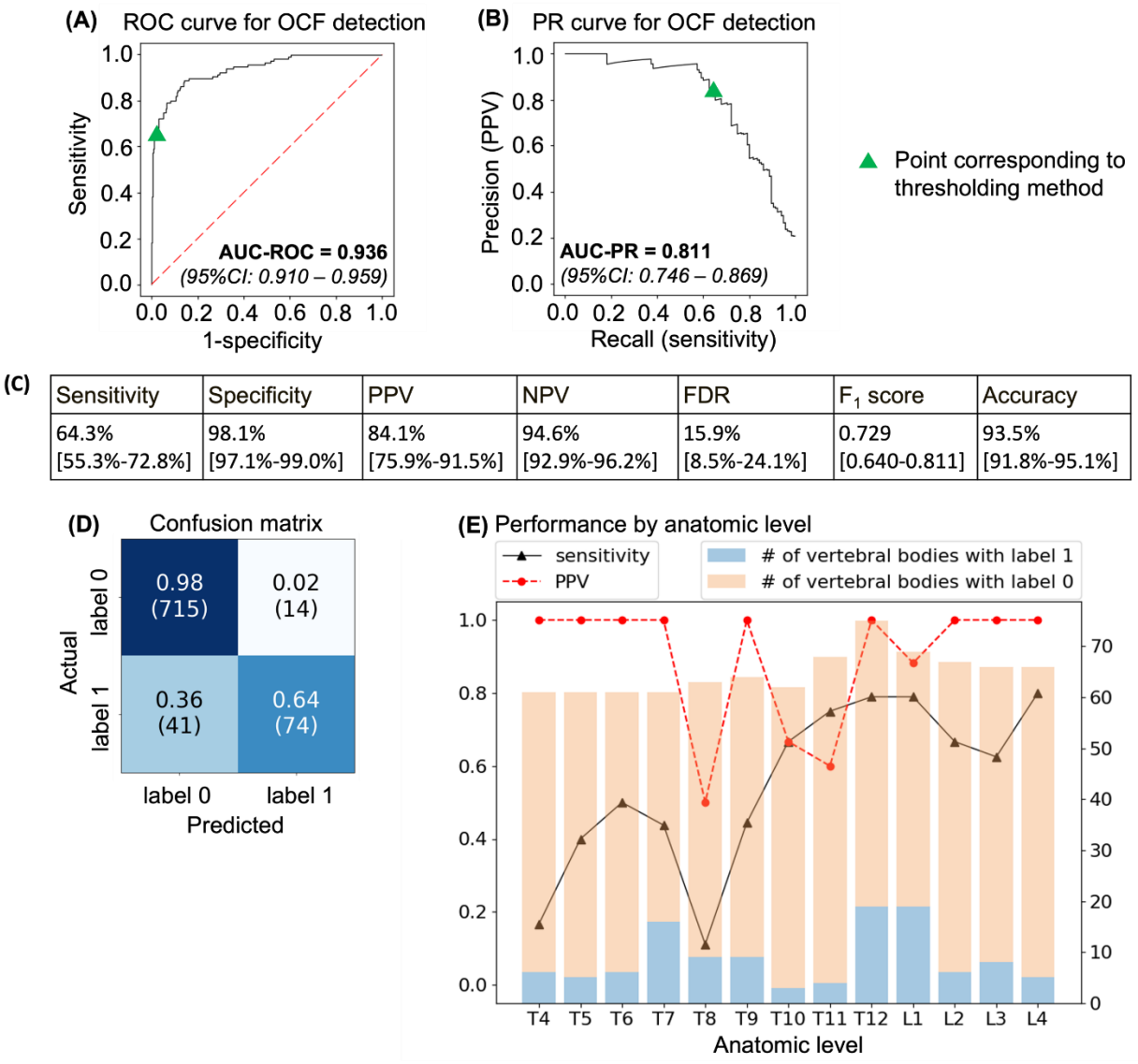
Supplementary Figure 6. Performance of the model built using the ensemble averaging algorithm in Task 1 and evaluated on the test set of the MrOS-m2ABQ dataset.



(C)

Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
41.1% [35.7%-46.8%]	99.8% [99.7%-99.9%]	91.0% [86.3%-95.6%]	97.3% [96.9%-97.7%]	9.0% [4.4%-13.7%]	0.566 [0.505-0.629]	97.2% [96.8%-97.6%]

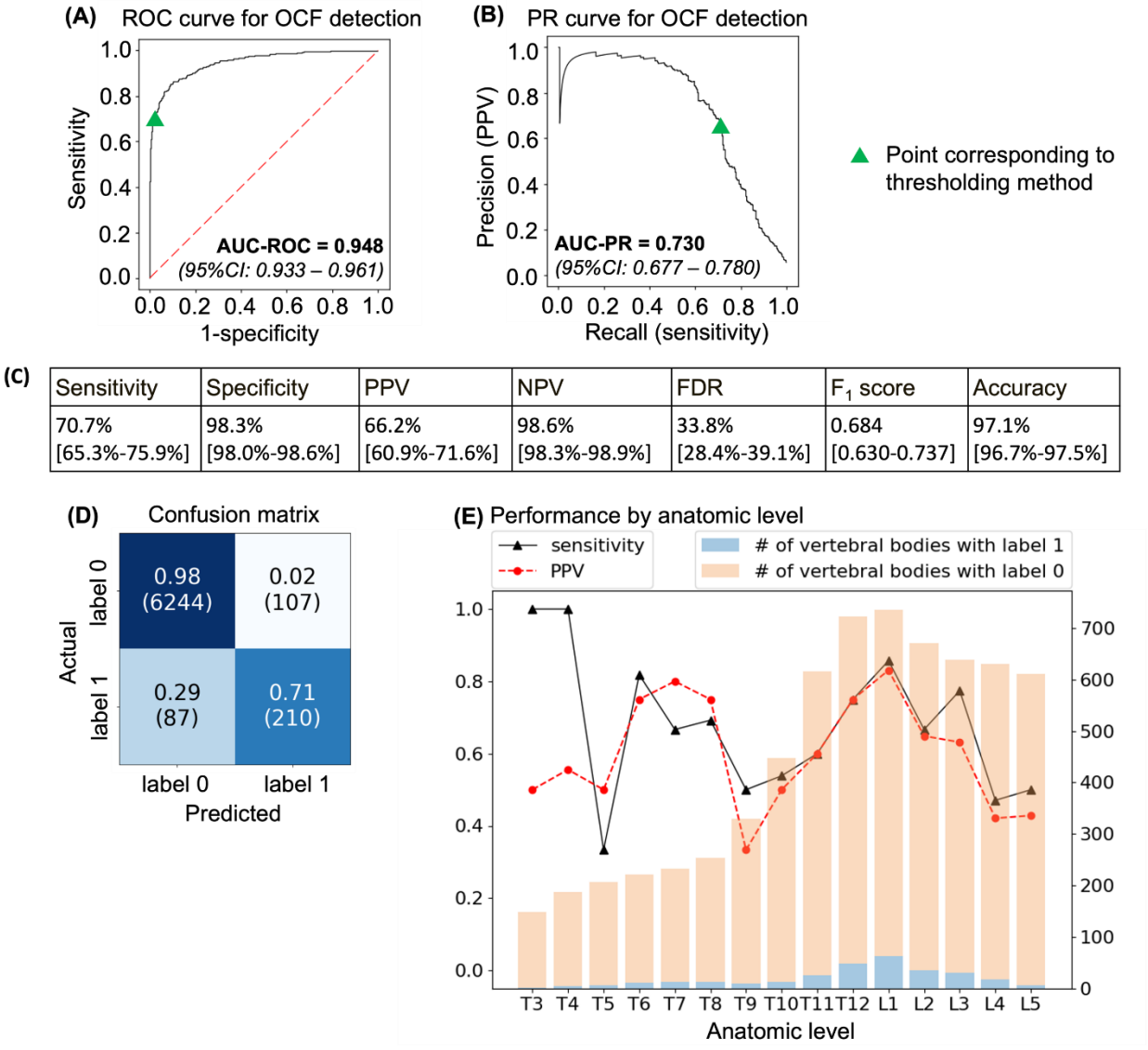
Supplementary Figure 7. Performance of the model built using the ensemble averaging algorithm in Task 1 and evaluated on the test set of the local-m2ABQ dataset.



(C)

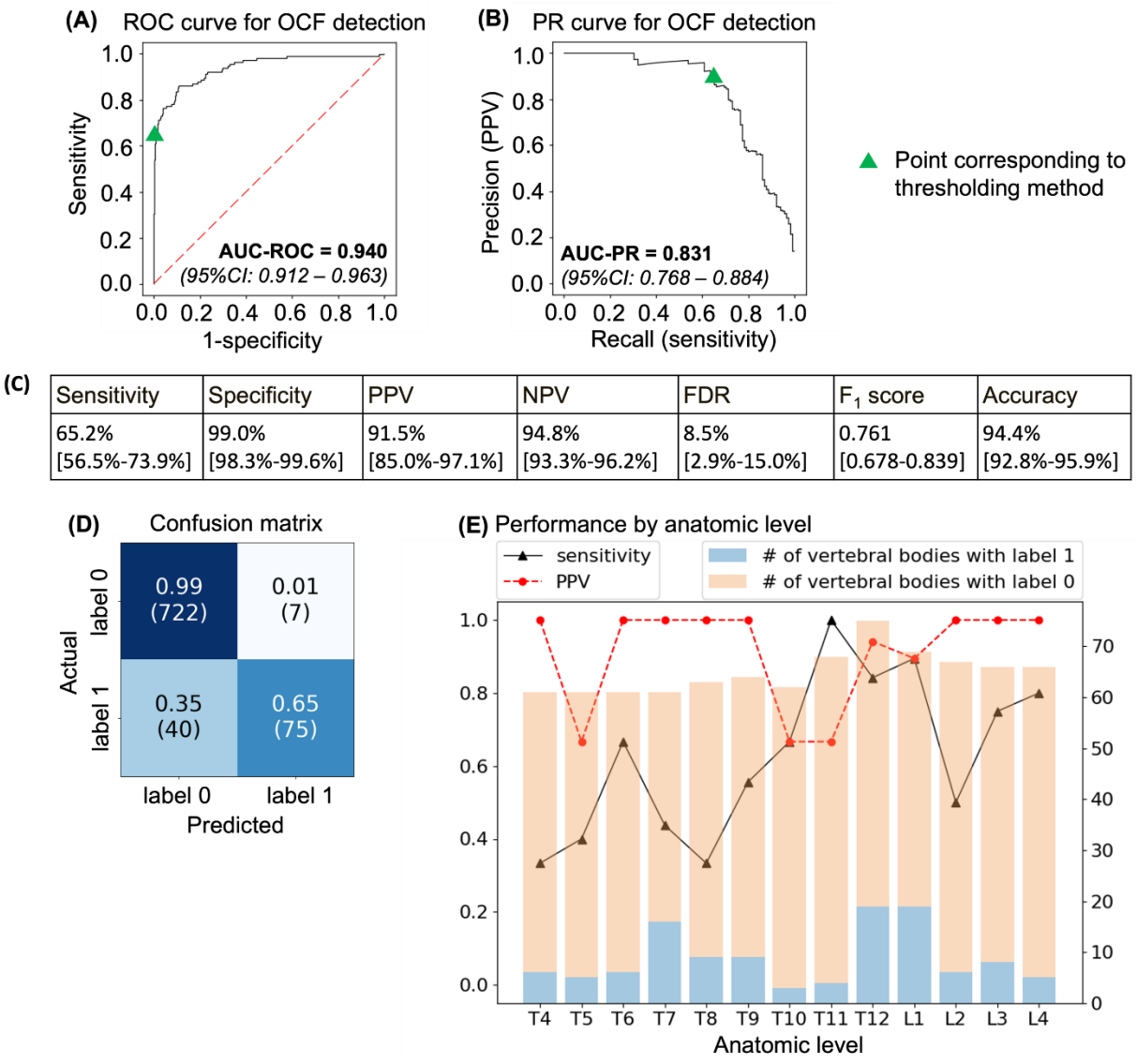
Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
64.3% [55.3%-72.8%]	98.1% [97.1%-99.0%]	84.1% [75.9%-91.5%]	94.6% [92.9%-96.2%]	15.9% [8.5%-24.1%]	0.729 [0.640-0.811]	93.5% [91.8%-95.1%]

Supplementary Figure 8. Performance of the model built using the ensemble averaging algorithm in Task 2 and evaluated on the test set of the MrOS-m2ABQ dataset.



Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
70.7% [65.3%-75.9%]	98.3% [98.0%-98.6%]	66.2% [60.9%-71.6%]	98.6% [98.3%-98.9%]	33.8% [28.4%-39.1%]	0.684 [0.630-0.737]	97.1% [96.7%-97.5%]

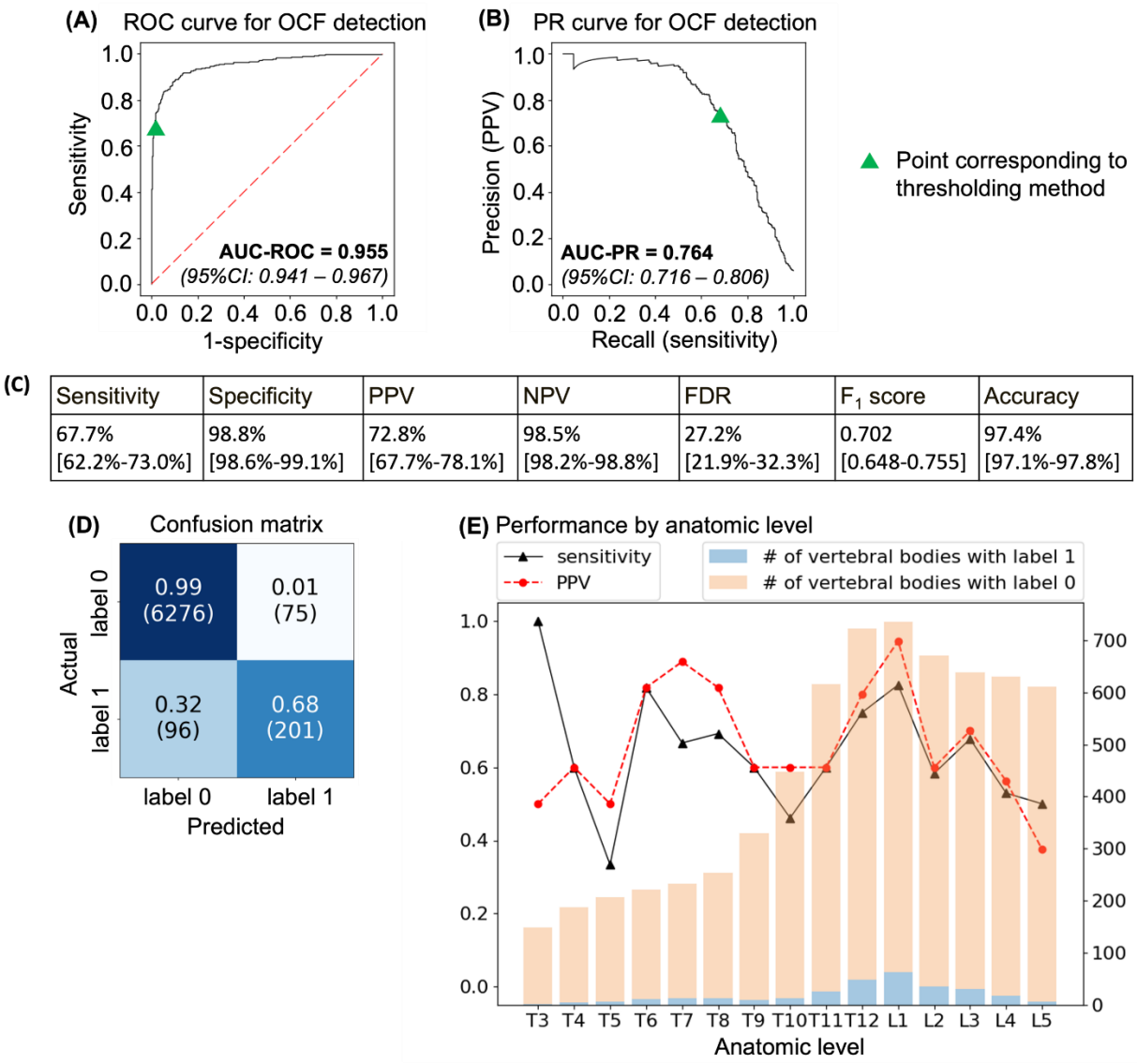
Supplementary Figure 9. Performance of the model built using the ensemble averaging algorithm in Task 2 and evaluated on the test set of the local-m2ABQ dataset.



(C)

Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
65.2% [56.5%-73.9%]	99.0% [98.3%-99.6%]	91.5% [85.0%-97.1%]	94.8% [93.3%-96.2%]	8.5% [2.9%-15.0%]	0.761 [0.678-0.839]	94.4% [92.8%-95.9%]

Supplementary Figure 10. Performance of the model built using the ensemble averaging algorithm in Task 3 and evaluated on the test set of the MrOS-m2ABQ dataset.

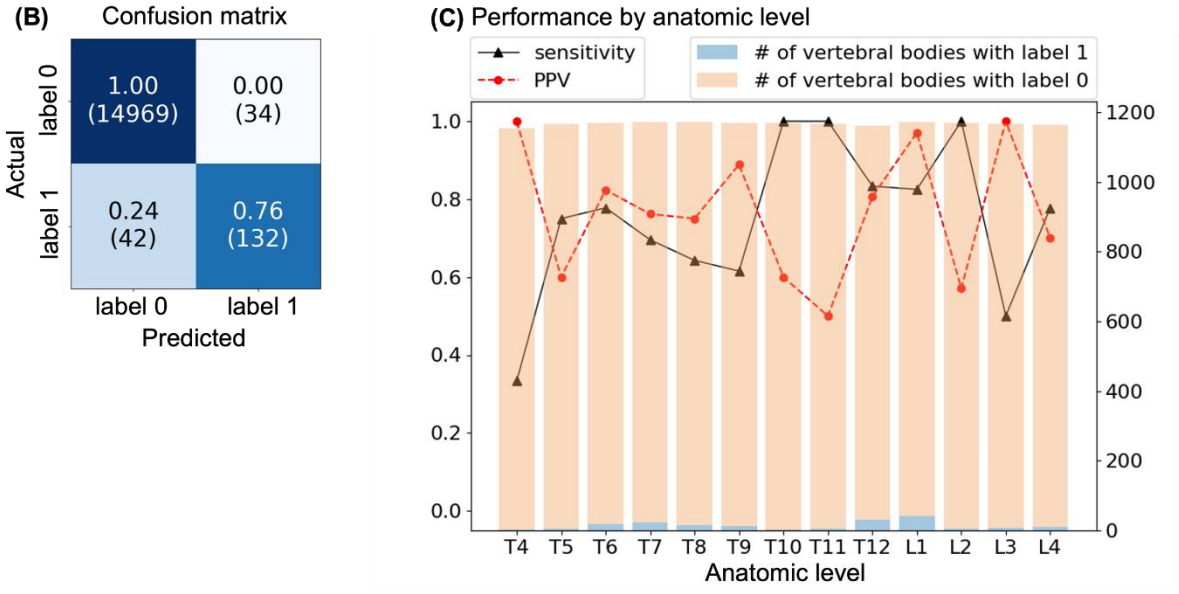


Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
67.7% [62.2%-73.0%]	98.8% [98.6%-99.1%]	72.8% [67.7%-78.1%]	98.5% [98.2%-98.8%]	27.2% [21.9%-32.3%]	0.702 [0.648-0.755]	97.4% [97.1%-97.8%]

Supplementary Figure 11. Performance of the model built using the ensemble averaging algorithm in Task 3 and evaluated on the test set of the local-m2ABQ dataset.

(A)

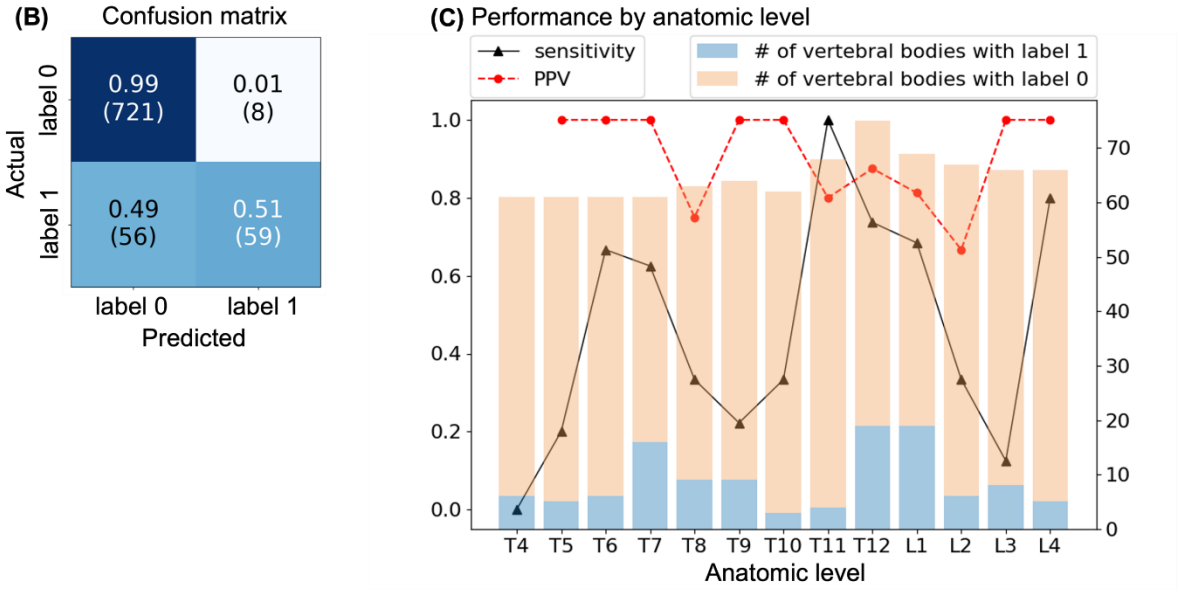
Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
75.9% [69.6%-82.5%]	99.8% [99.7%-99.8%]	79.5% [73.0%-85.6%]	99.7% [99.6%-99.8%]	20.5% [14.4%-27.0%]	0.776 [0.712-0.840]	99.5% [99.4%-99.6%]



Supplementary Figure 12. Performance of the model built using the ensemble majority voting algorithm in Task 1 and evaluated on the test set of the MrOS-mSQ dataset.

(A)

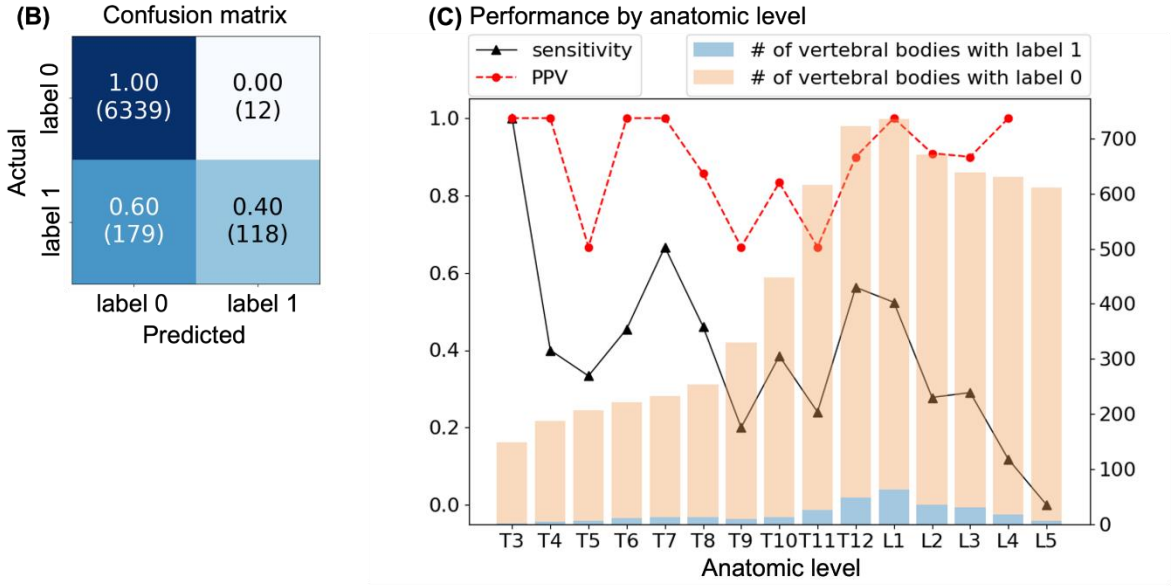
Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
51.3% [41.8%-60.7%]	98.9% [98.1%-99.6%]	88.1% [80.0%-95.5%]	92.8% [90.9%-94.6%]	11.9% [4.5%-20.0%]	0.648 [0.549-0.742]	92.4% [90.6%-94.2%]



Supplementary Figure 13. Performance of the model built using the ensemble majority voting algorithm in Task 1 and evaluated on the test set of the MrOS-m2ABQ dataset.

(A)

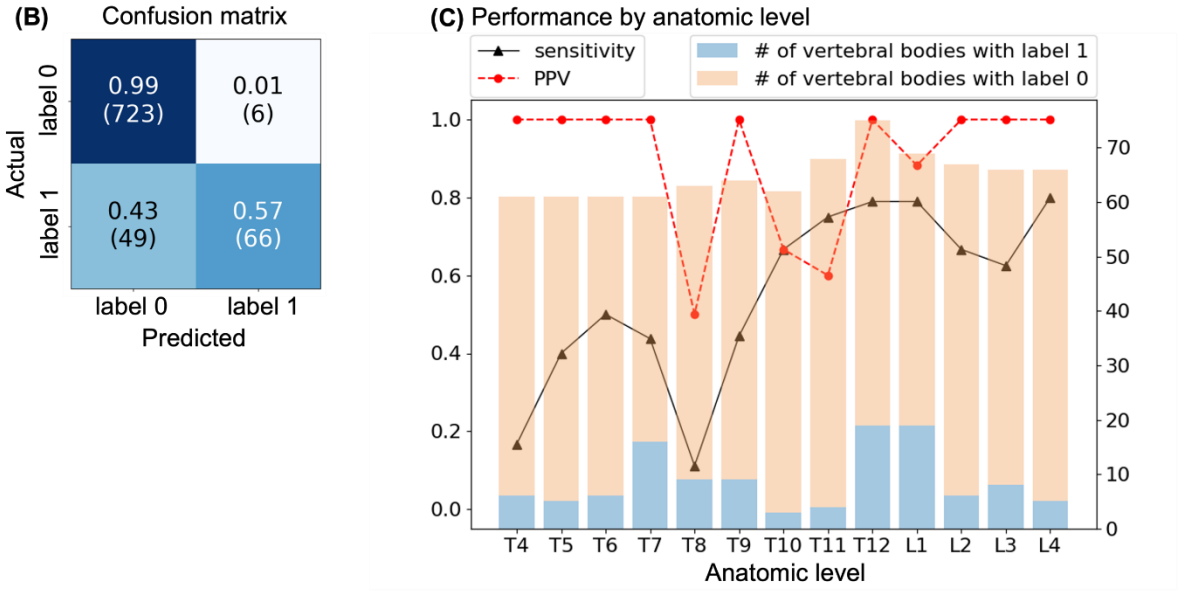
Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
39.7% [34.4%-45.2%]	99.8% [99.7%-99.9%]	90.8% [85.4%-95.5%]	97.3% [96.8%-97.6%]	9.2% [4.5%-14.6%]	0.553 [0.491-0.614]	97.1% [96.7%-97.5%]



Supplementary Figure 14. Performance of the model built using the ensemble majority voting algorithm in Task 1 and evaluated on the test set of the local-m2ABQ dataset.

(A)

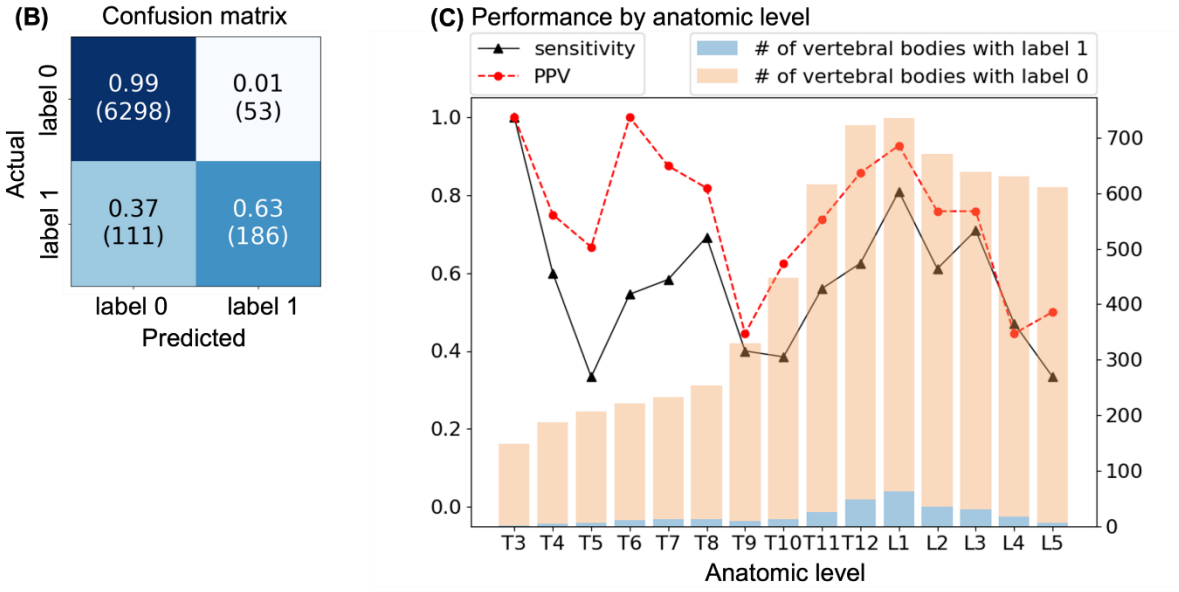
Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
57.4% [48.5%-66.1%]	99.2% [98.5%-99.7%]	91.7% [84.6%-97.3%]	93.7% [91.9%-95.3%]	8.3% [2.7%-15.4%]	0.706 [0.616-0.787]	93.5% [91.7%-95.1%]



Supplementary Figure 15. Performance of the model built using the ensemble majority voting algorithm in Task 2 and evaluated on the test set of the MrOS-m2ABQ dataset.

(A)

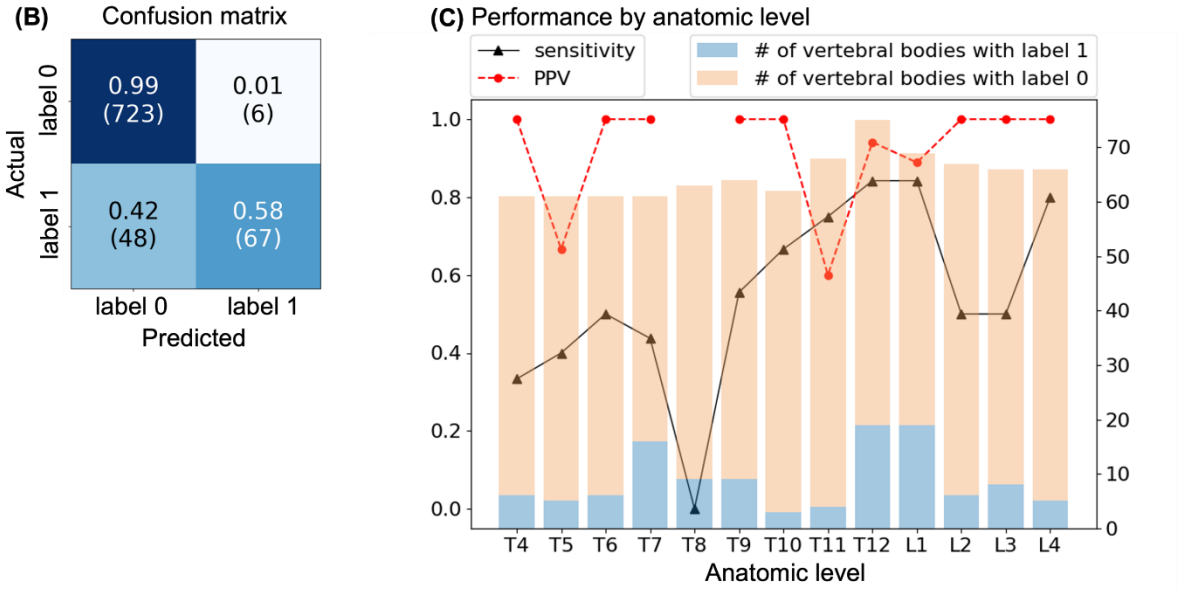
Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
62.6% [57.3%-68.0%]	99.2% [98.9%-99.4%]	77.8% [72.5%-83.1%]	98.3% [97.9%-98.6%]	22.2% [16.9%-27.5%]	0.694 [0.640-0.748]	97.5% [97.1%-97.9%]



Supplementary Figure 16. Performance of the model built using the ensemble majority voting algorithm in Task 2 and evaluated on the test set of the local-m2ABQ dataset.

(A)

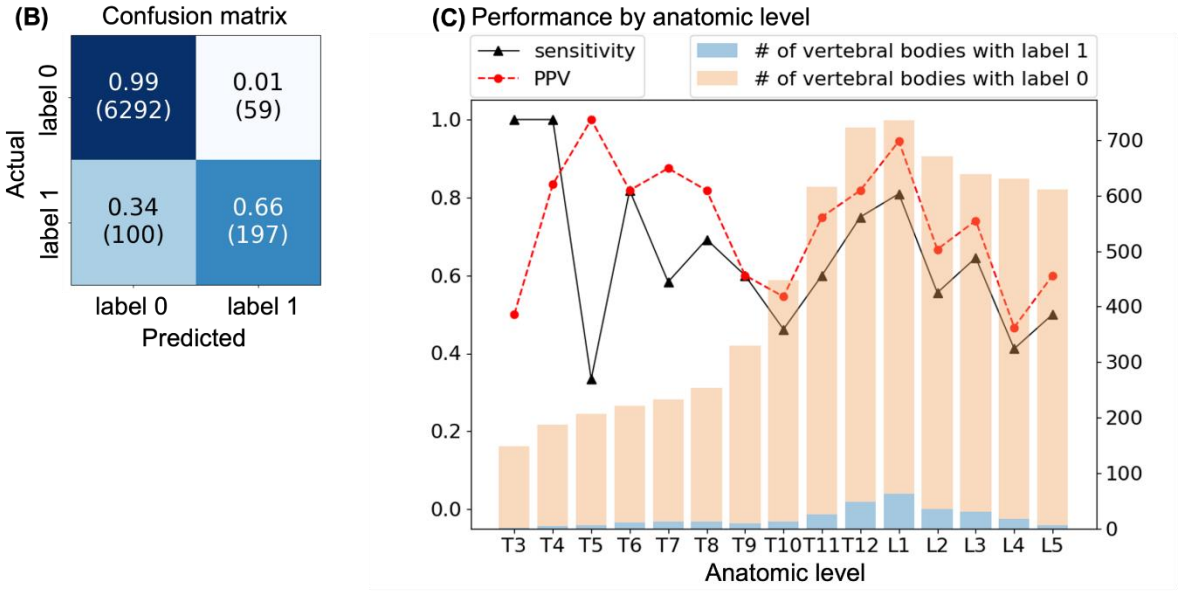
Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
58.3% [49.6%-67.2%]	99.2% [98.5%-99.7%]	91.8% [84.6%-97.3%]	93.8% [92.1%-95.4%]	8.2% [2.7%-15.4%]	0.713 [0.625-0.795]	93.6% [91.8%-95.1%]



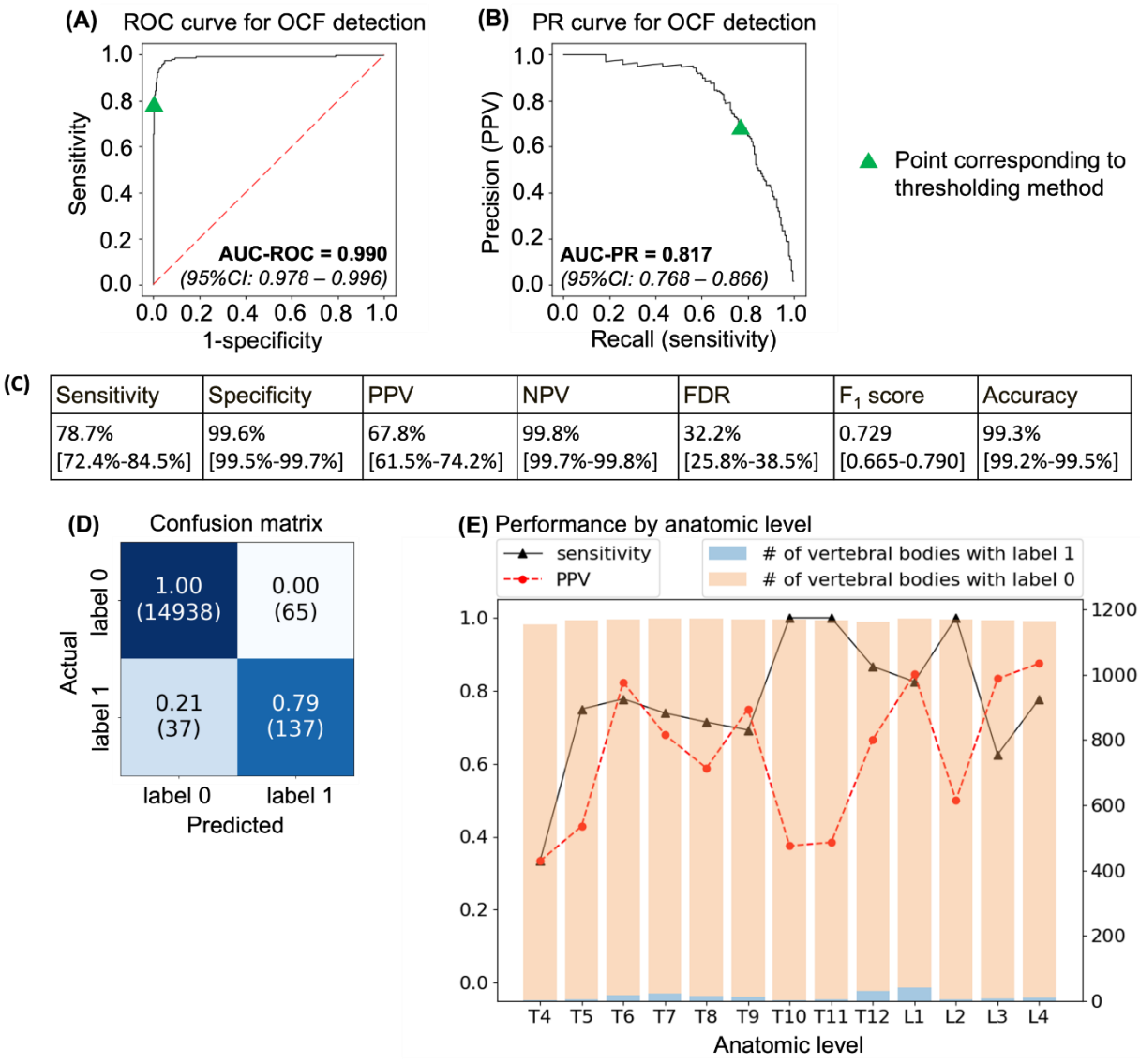
Supplementary Figure 17. Performance of the model built using the ensemble majority voting algorithm in Task 3 and evaluated on the test set of the MrOS-m2ABQ dataset.

(A)

Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
66.3% [60.7%-71.5%]	99.1% [98.8%-99.3%]	77.0% [71.7%-82.2%]	98.4% [98.1%-98.8%]	23.0% [17.8%-28.3%]	0.712 [0.657-0.765]	97.6% [97.2%-98.0%]



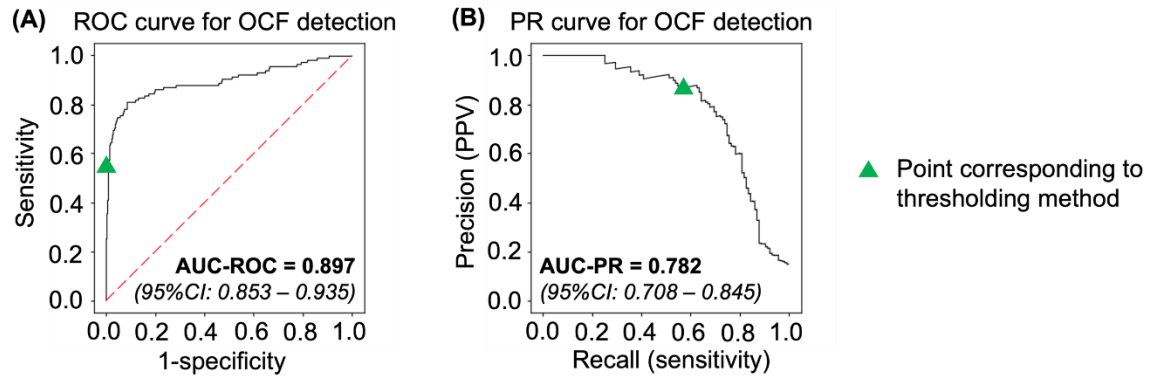
Supplementary Figure 18. Performance of the model built using the ensemble majority voting algorithm in Task 3 and evaluated on the test set of the local-m2ABQ dataset.



(C)

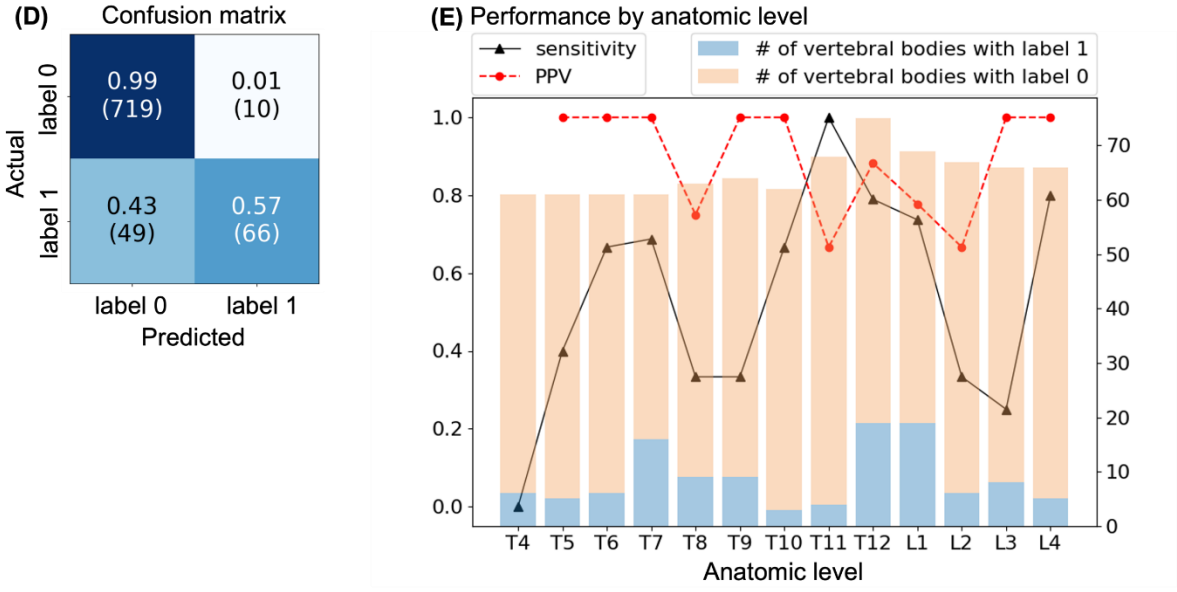
Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
78.7% [72.4%-84.5%]	99.6% [99.5%-99.7%]	67.8% [61.5%-74.2%]	99.8% [99.7%-99.8%]	32.2% [25.8%-38.5%]	0.729 [0.665-0.790]	99.3% [99.2%-99.5%]

Supplementary Figure 19. Performance of the model built using GoogLeNet in Task 1 and evaluated on the test set of the MrOS-mSQ dataset.

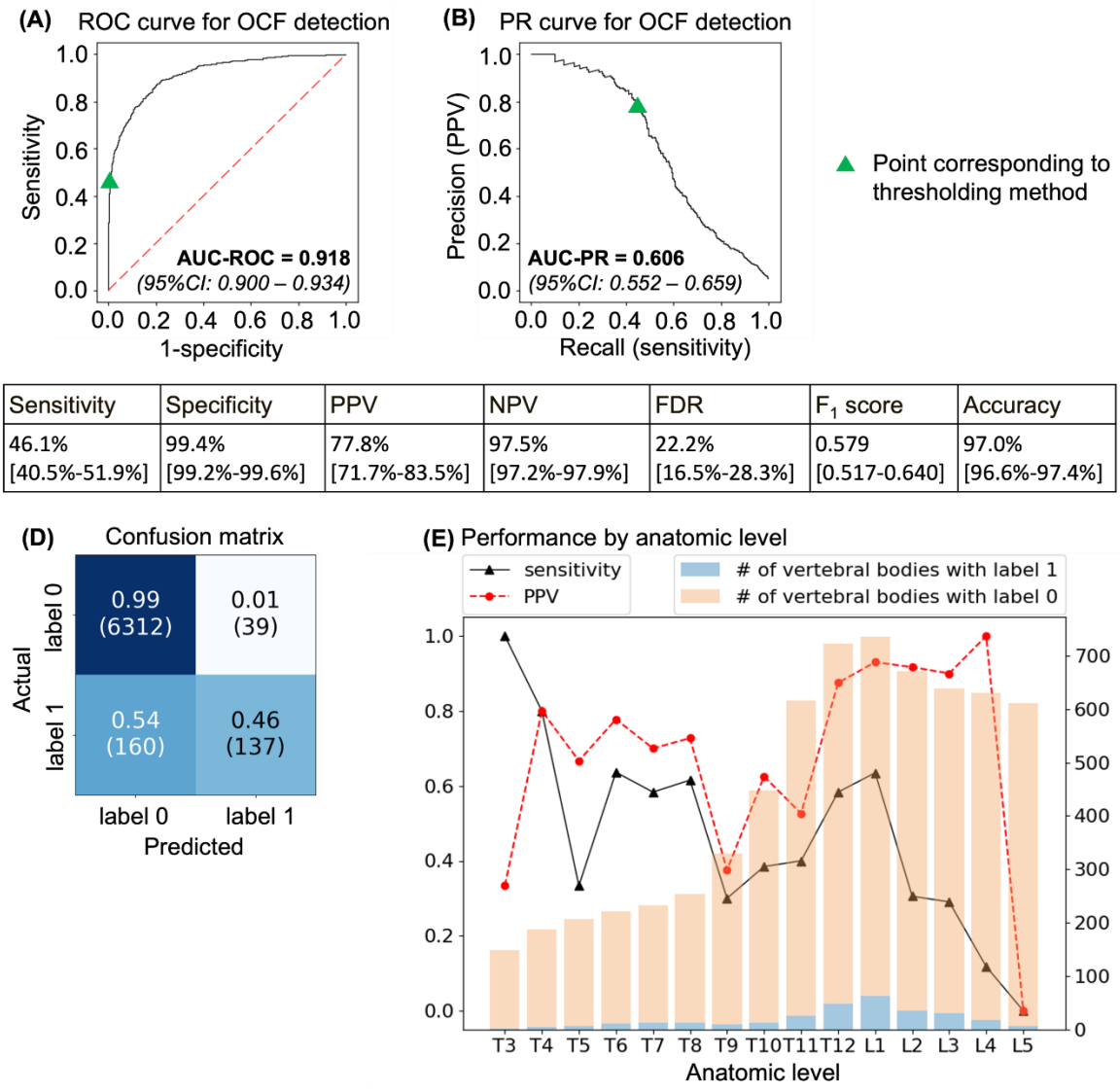


(C)

Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
57.4% [48.2%-66.4%]	98.6% [97.8%-99.4%]	86.8% [78.3%-94.1%]	93.6% [91.8%-95.2%]	13.2% [5.9%-21.7%]	0.691 [0.596-0.778]	93.0% [91.2%-94.5%]

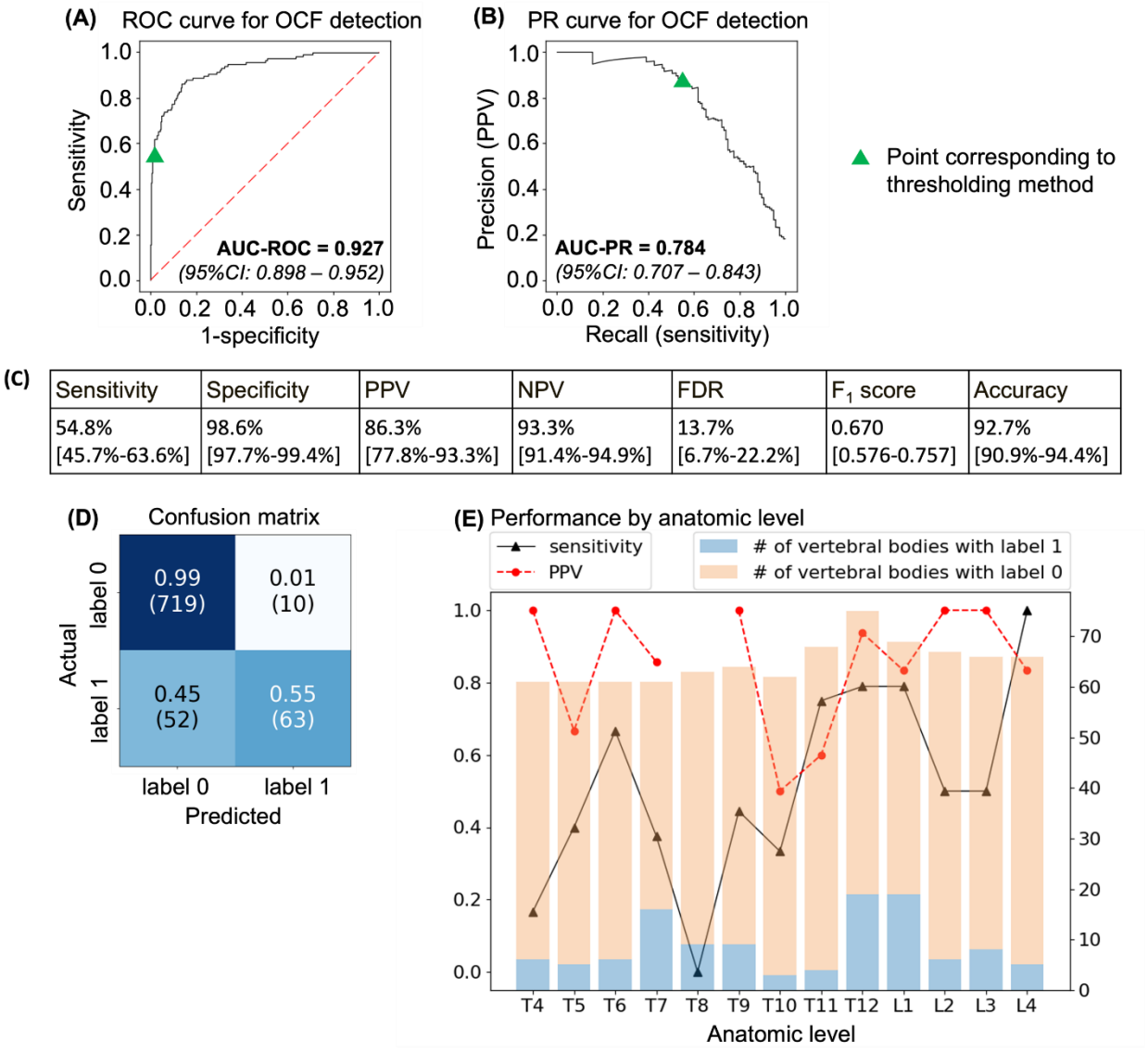


Supplementary Figure 20. Performance of the model built using GoogLeNet in Task 1 and evaluated on the test set of the MrOS-m2ABQ dataset.



Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
46.1% [40.5%-51.9%]	99.4% [99.2%-99.6%]	77.8% [71.7%-83.5%]	97.5% [97.2%-97.9%]	22.2% [16.5%-28.3%]	0.579 [0.517-0.640]	97.0% [96.6%-97.4%]

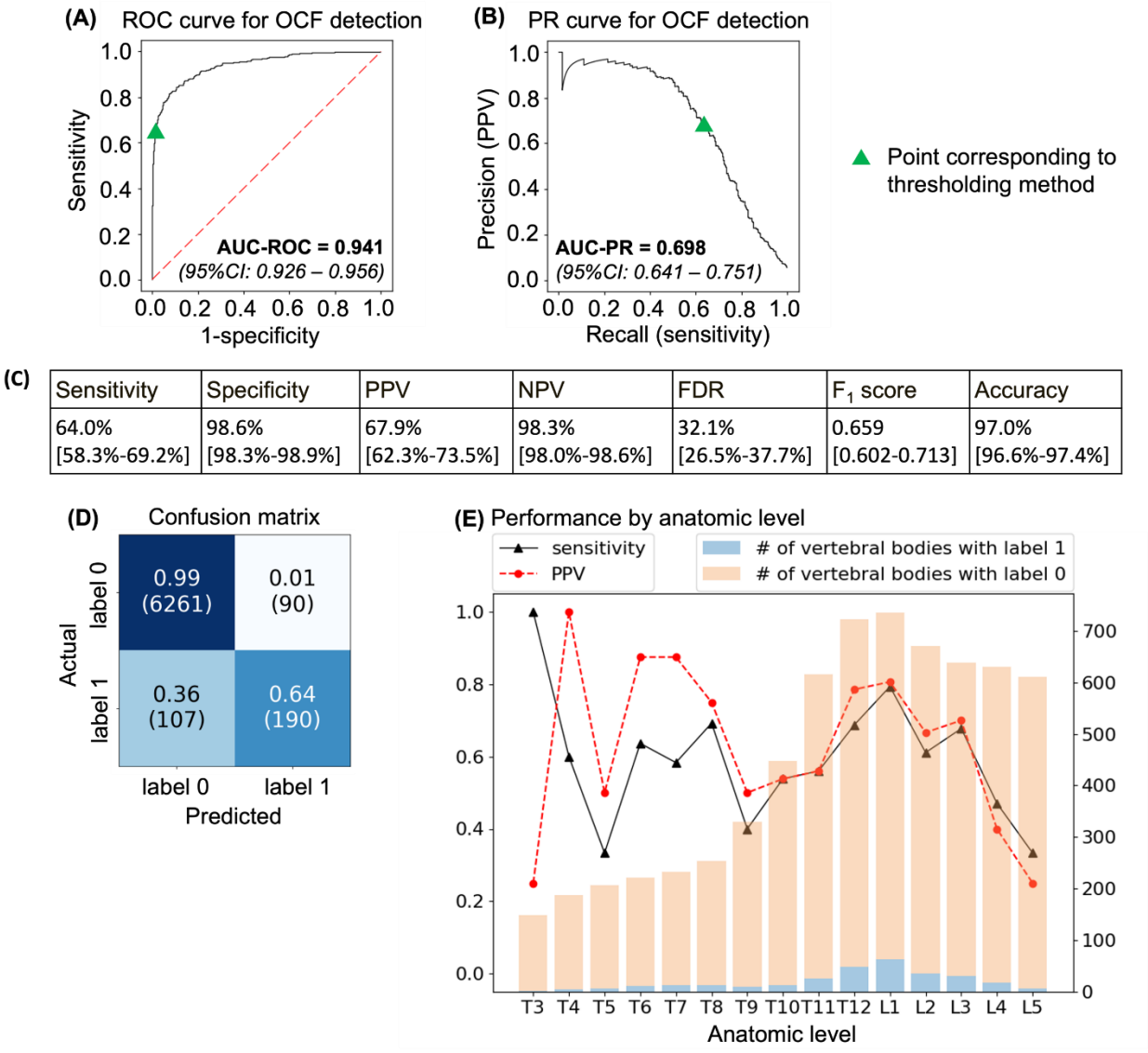
Supplementary Figure 21. Performance of the model built using GoogLeNet in Task 1 and evaluated on the test set of the local-m2ABQ dataset.



(C)

Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
54.8% [45.7%-63.6%]	98.6% [97.7%-99.4%]	86.3% [77.8%-93.3%]	93.3% [91.4%-94.9%]	13.7% [6.7%-22.2%]	0.670 [0.576-0.757]	92.7% [90.9%-94.4%]

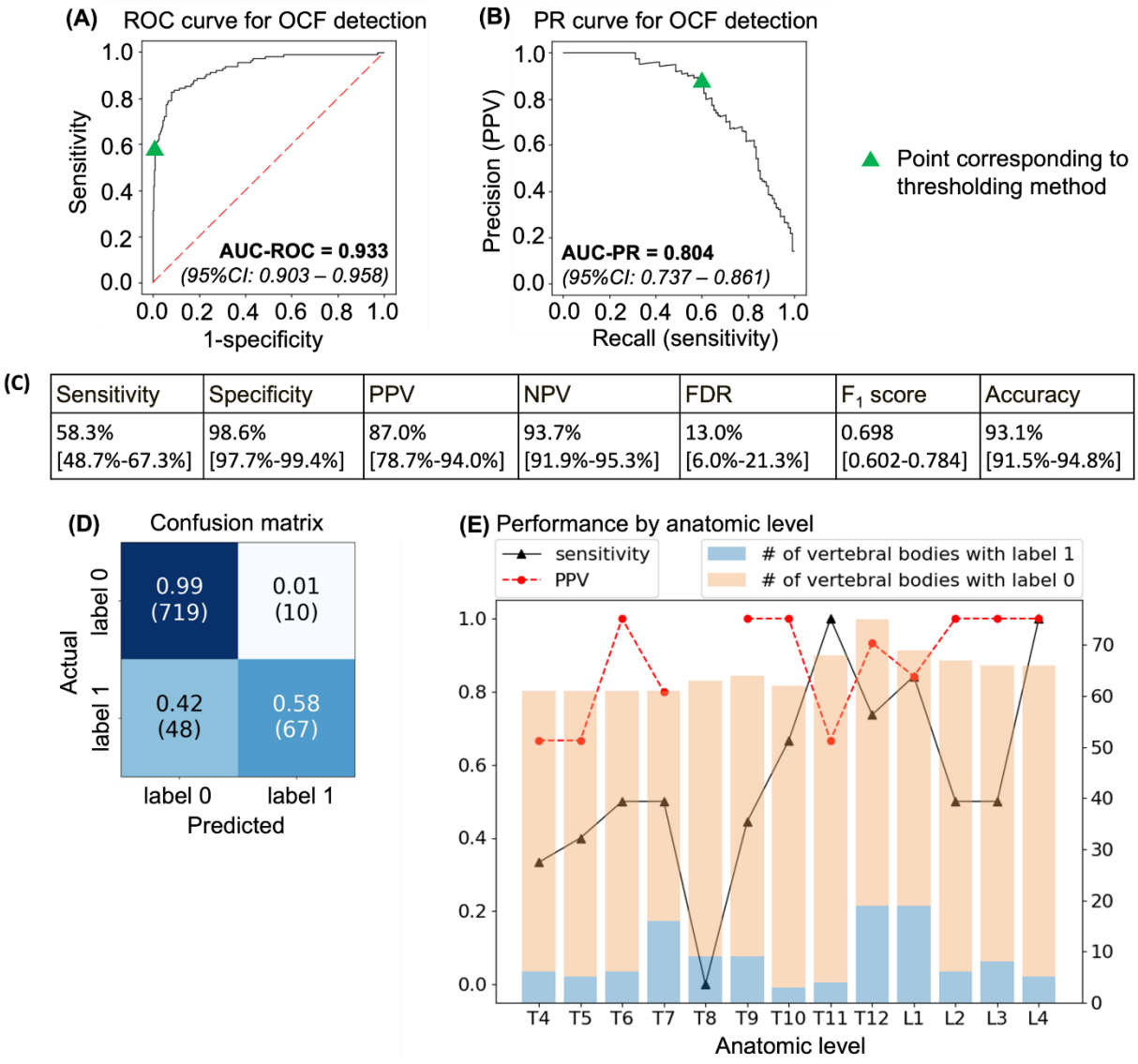
Supplementary Figure 22. Performance of the model built using GoogLeNet in Task 2 and evaluated on the test set of the MrOS-m2ABQ dataset.



(C)

Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
64.0% [58.3%-69.2%]	98.6% [98.3%-98.9%]	67.9% [62.3%-73.5%]	98.3% [98.0%-98.6%]	32.1% [26.5%-37.7%]	0.659 [0.602-0.713]	97.0% [96.6%-97.4%]

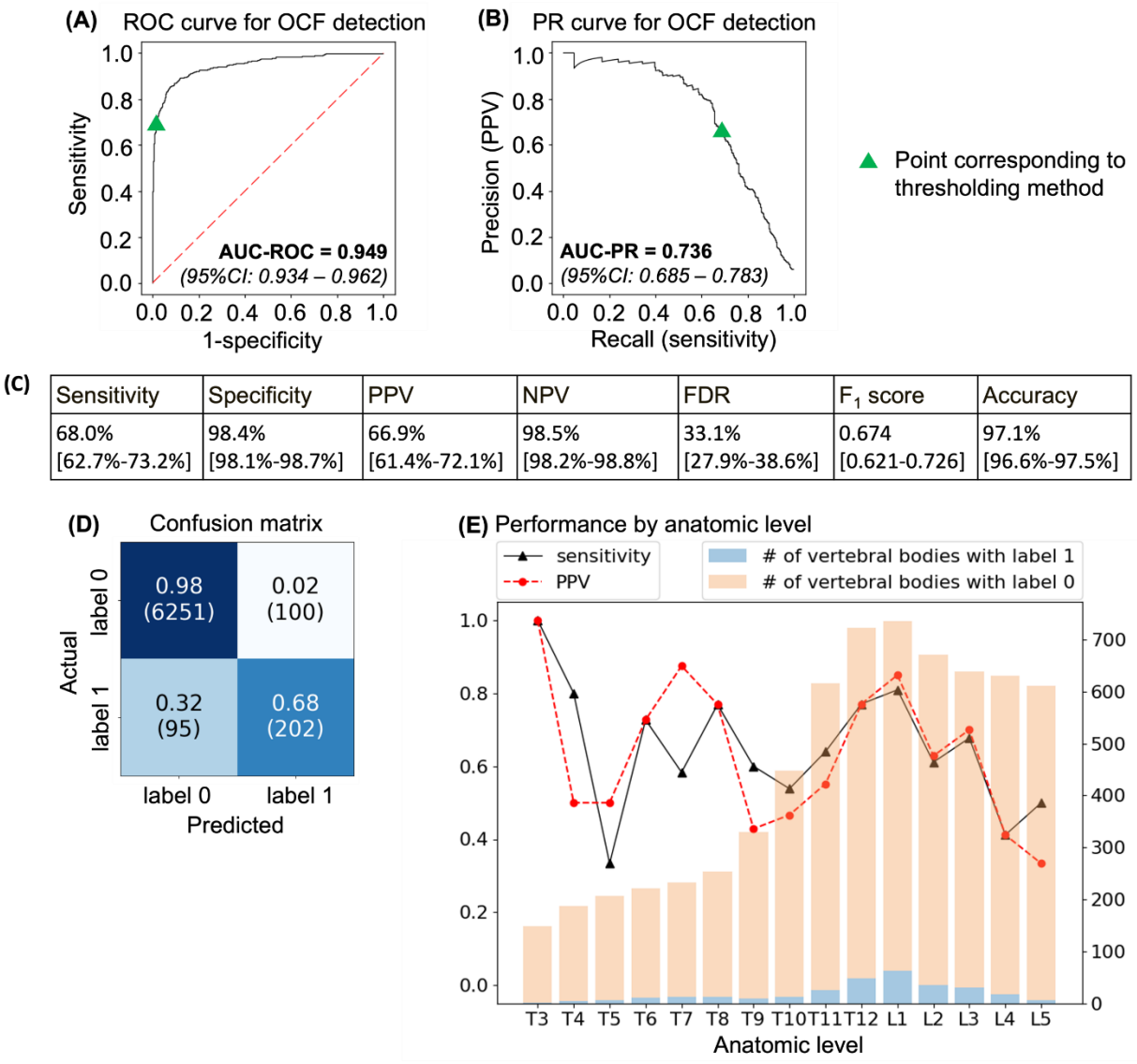
Supplementary Figure 23. Performance of the model built using GoogLeNet in Task 2 and evaluated on the test set of the local-m2ABQ dataset.



(C)

Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
58.3% [48.7%-67.3%]	98.6% [97.7%-99.4%]	87.0% [78.7%-94.0%]	93.7% [91.9%-95.3%]	13.0% [6.0%-21.3%]	0.698 [0.602-0.784]	93.1% [91.5%-94.8%]

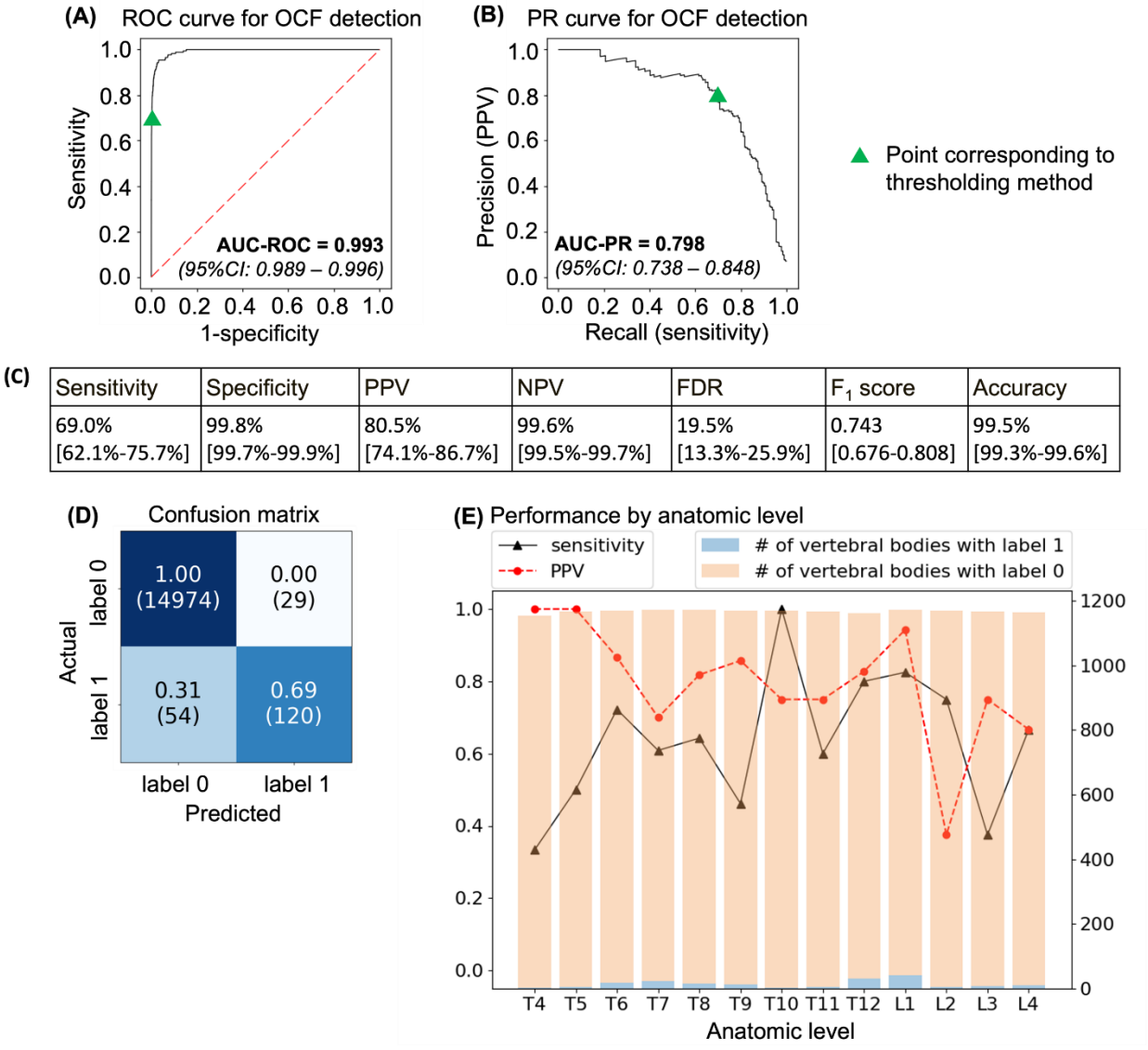
Supplementary Figure 24. Performance of the model built using GoogLeNet in Task 3 and evaluated on the test set of the MrOS-m2ABQ dataset.



(C)

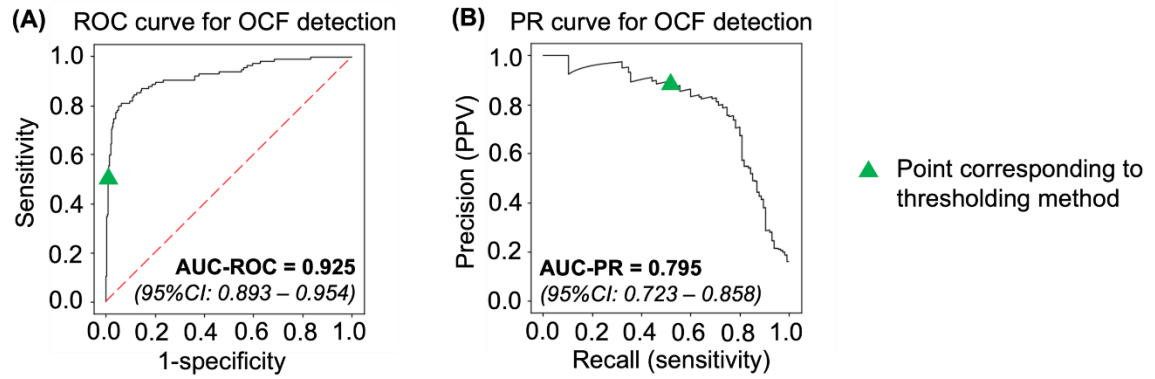
Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
68.0% [62.7%-73.2%]	98.4% [98.1%-98.7%]	66.9% [61.4%-72.1%]	98.5% [98.2%-98.8%]	33.1% [27.9%-38.6%]	0.674 [0.621-0.726]	97.1% [96.6%-97.5%]

Supplementary Figure 25. Performance of the model built using GoogLeNet in Task 3 and evaluated on the test set of the local-m2ABQ dataset.



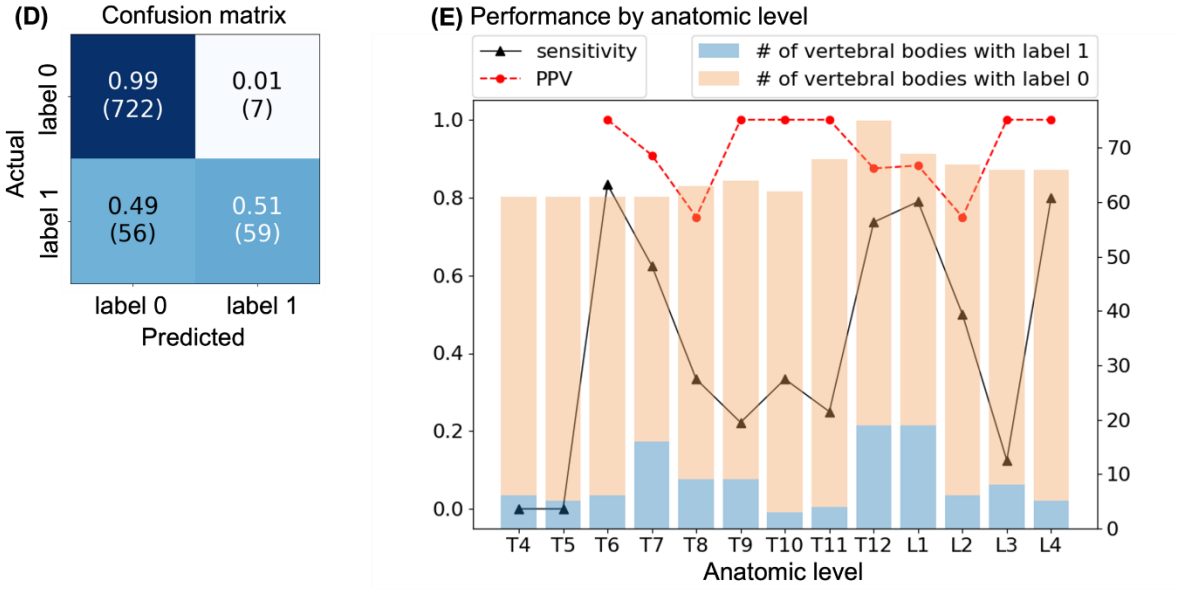
Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
69.0% [62.1%-75.7%]	99.8% [99.7%-99.9%]	80.5% [74.1%-86.7%]	99.6% [99.5%-99.7%]	19.5% [13.3%-25.9%]	0.743 [0.676-0.808]	99.5% [99.3%-99.6%]

Supplementary Figure 26. Performance of the model built using Inception-ResNet-v2 in Task 1 and evaluated on the test set of the MrOS-mSQ dataset.

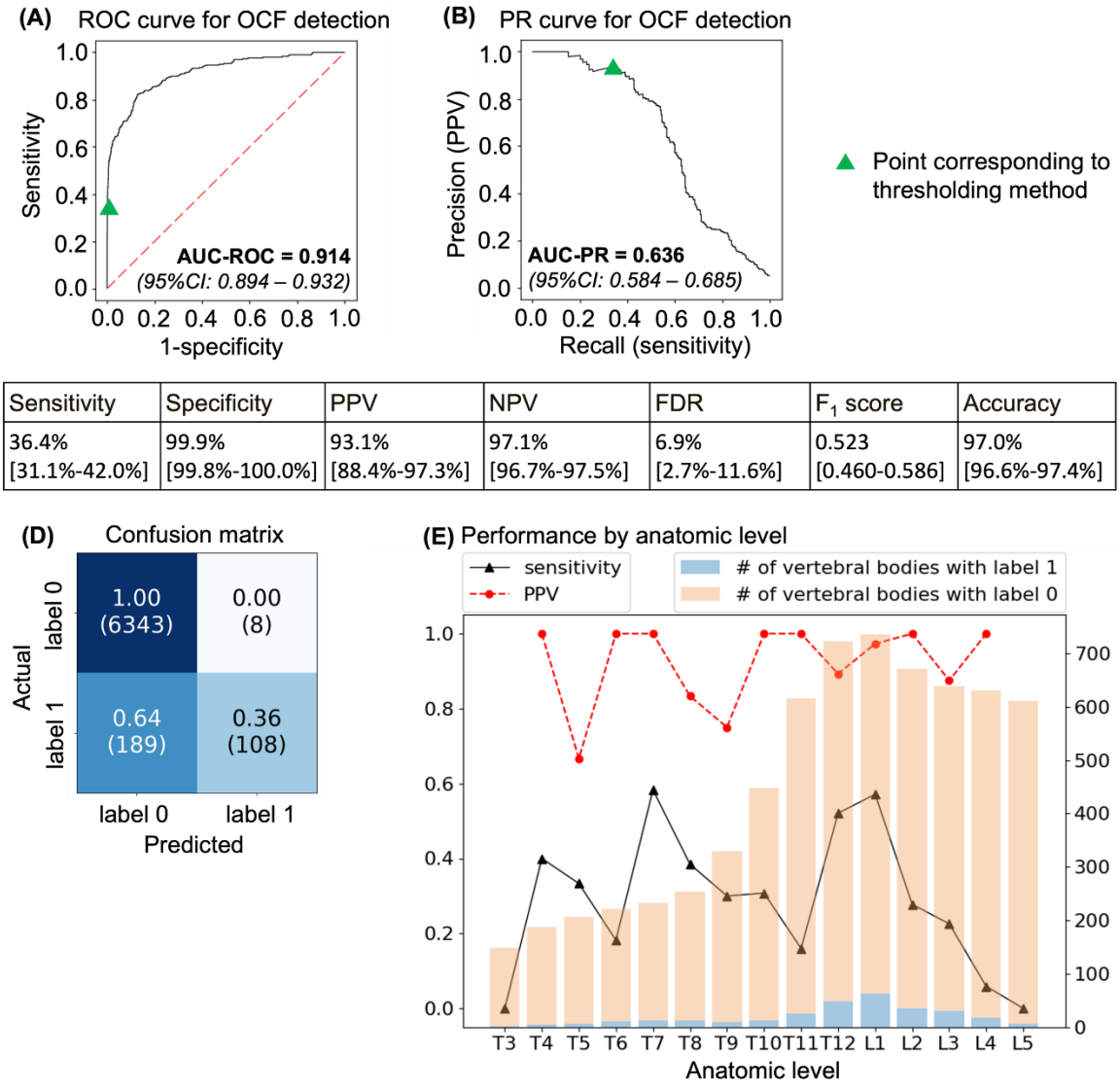


(C)

Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
51.3% [42.1%-60.6%]	99.0% [98.2%-99.6%]	89.4% [81.8%-96.3%]	92.8% [91.0%-94.6%]	10.6% [3.7%-18.2%]	0.652 [0.556-0.744]	92.5% [90.8%-94.3%]



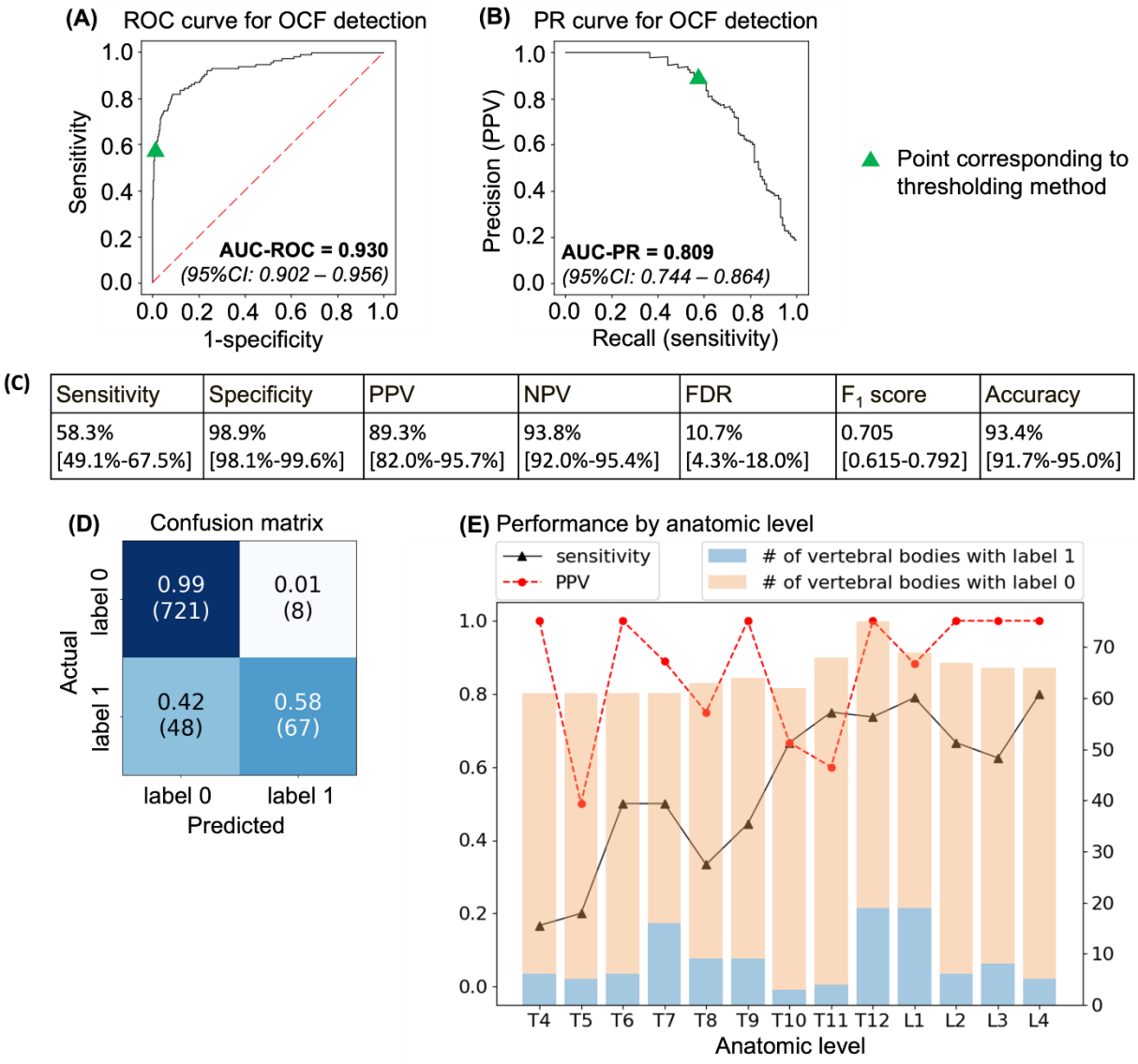
Supplementary Figure 27. Performance of the model built using Inception-ResNet-v2 in Task 1 and evaluated on the test set of the MrOS-m2ABQ dataset.



(C)

Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
36.4% [31.1%-42.0%]	99.9% [99.8%-100.0%]	93.1% [88.4%-97.3%]	97.1% [96.7%-97.5%]	6.9% [2.7%-11.6%]	0.523 [0.460-0.586]	97.0% [96.6%-97.4%]

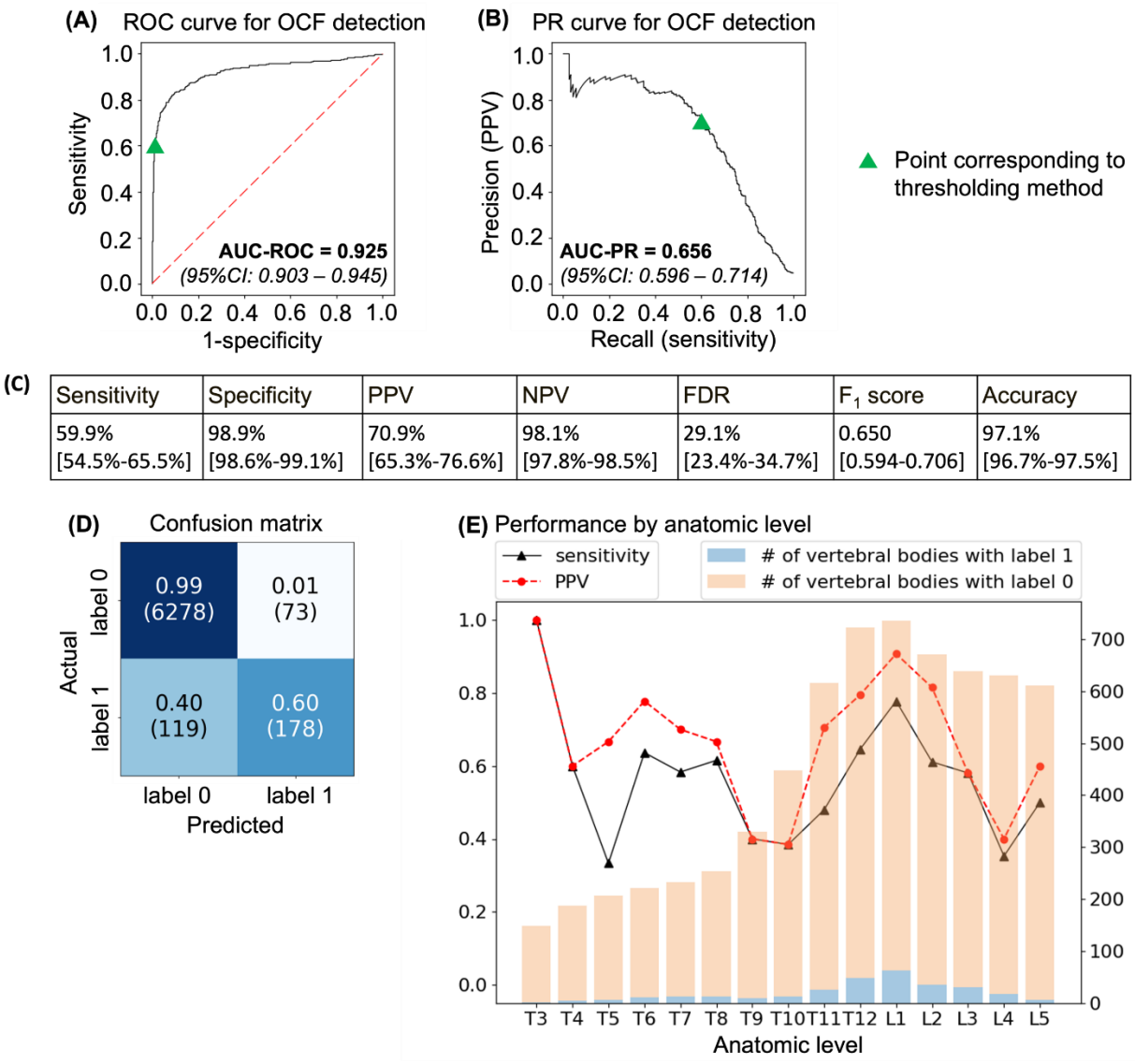
Supplementary Figure 28. Performance of the model built using Inception-ResNet-v2 in Task 1 and evaluated on the test set of the local-m2ABQ dataset.



(C)

Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
58.3% [49.1%-67.5%]	98.9% [98.1%-99.6%]	89.3% [82.0%-95.7%]	93.8% [92.0%-95.4%]	10.7% [4.3%-18.0%]	0.705 [0.615-0.792]	93.4% [91.7%-95.0%]

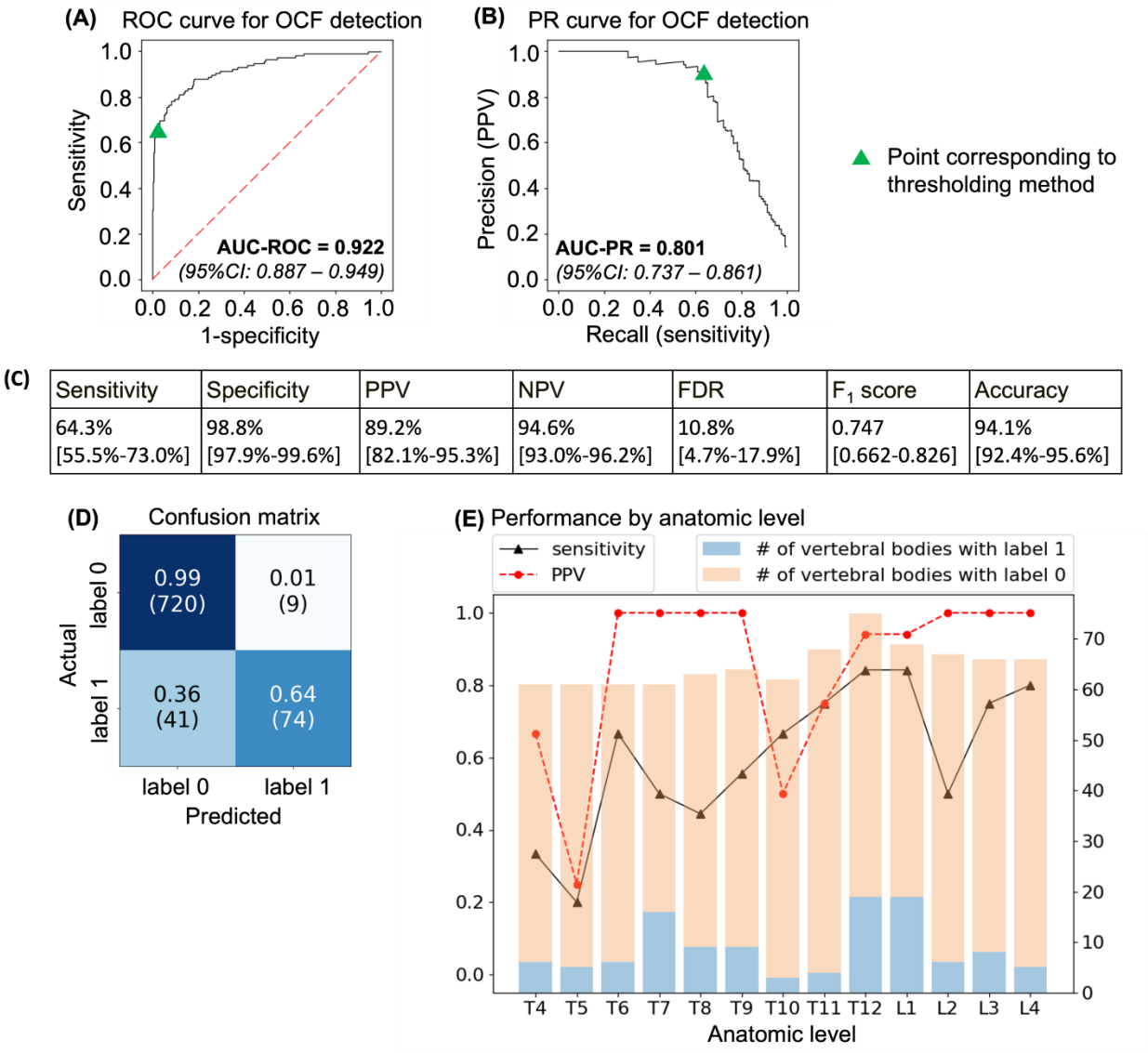
Supplementary Figure 29. Performance of the model built using Inception-ResNet-v2 in Task 2 and evaluated on the test set of the MrOS-m2ABQ dataset.



(C)

Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
59.9% [54.5%-65.5%]	98.9% [98.6%-99.1%]	70.9% [65.3%-76.6%]	98.1% [97.8%-98.5%]	29.1% [23.4%-34.7%]	0.650 [0.594-0.706]	97.1% [96.7%-97.5%]

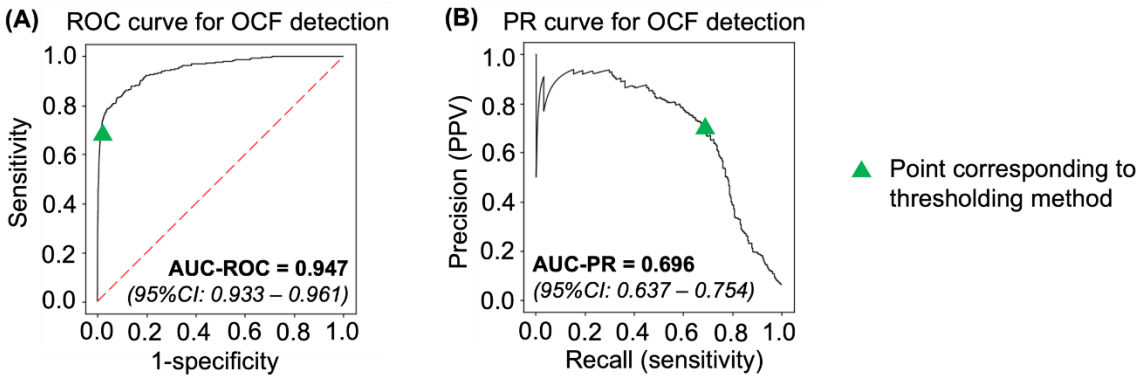
Supplementary Figure 30. Performance of the model built using Inception-ResNet-v2 in Task 2 and evaluated on the test set of the local-m2ABQ dataset.



(C)

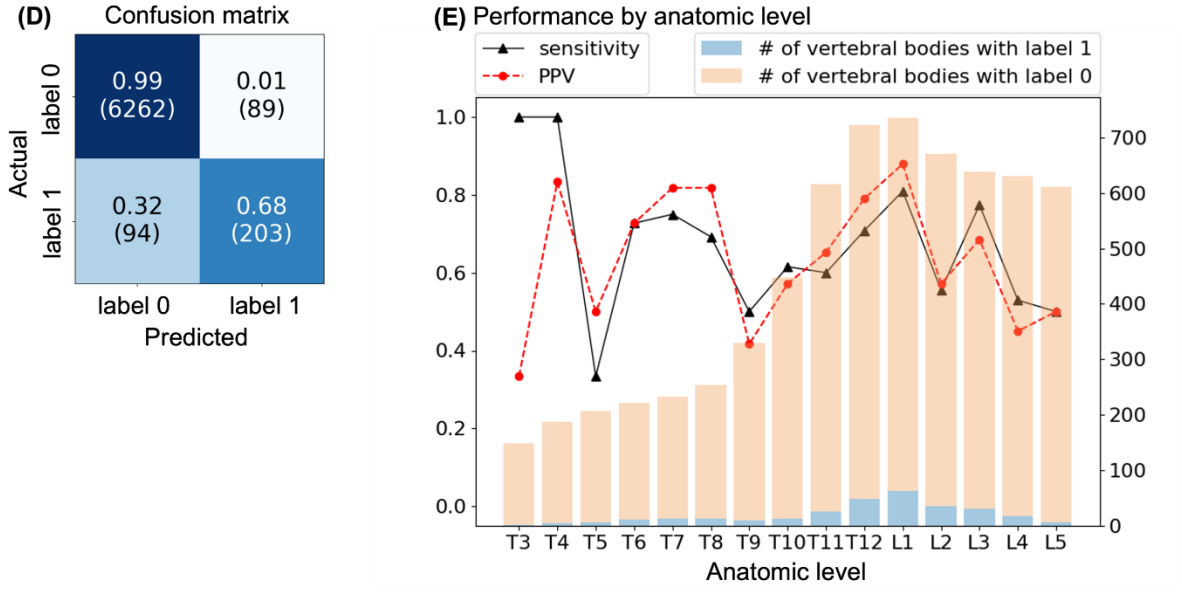
Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
64.3% [55.5%-73.0%]	98.8% [97.9%-99.6%]	89.2% [82.1%-95.3%]	94.6% [93.0%-96.2%]	10.8% [4.7%-17.9%]	0.747 [0.662-0.826]	94.1% [92.4%-95.6%]

Supplementary Figure 31. Performance of the model built using Inception-ResNet-v2 in Task 3 and evaluated on the test set of the MrOS-m2ABQ dataset.

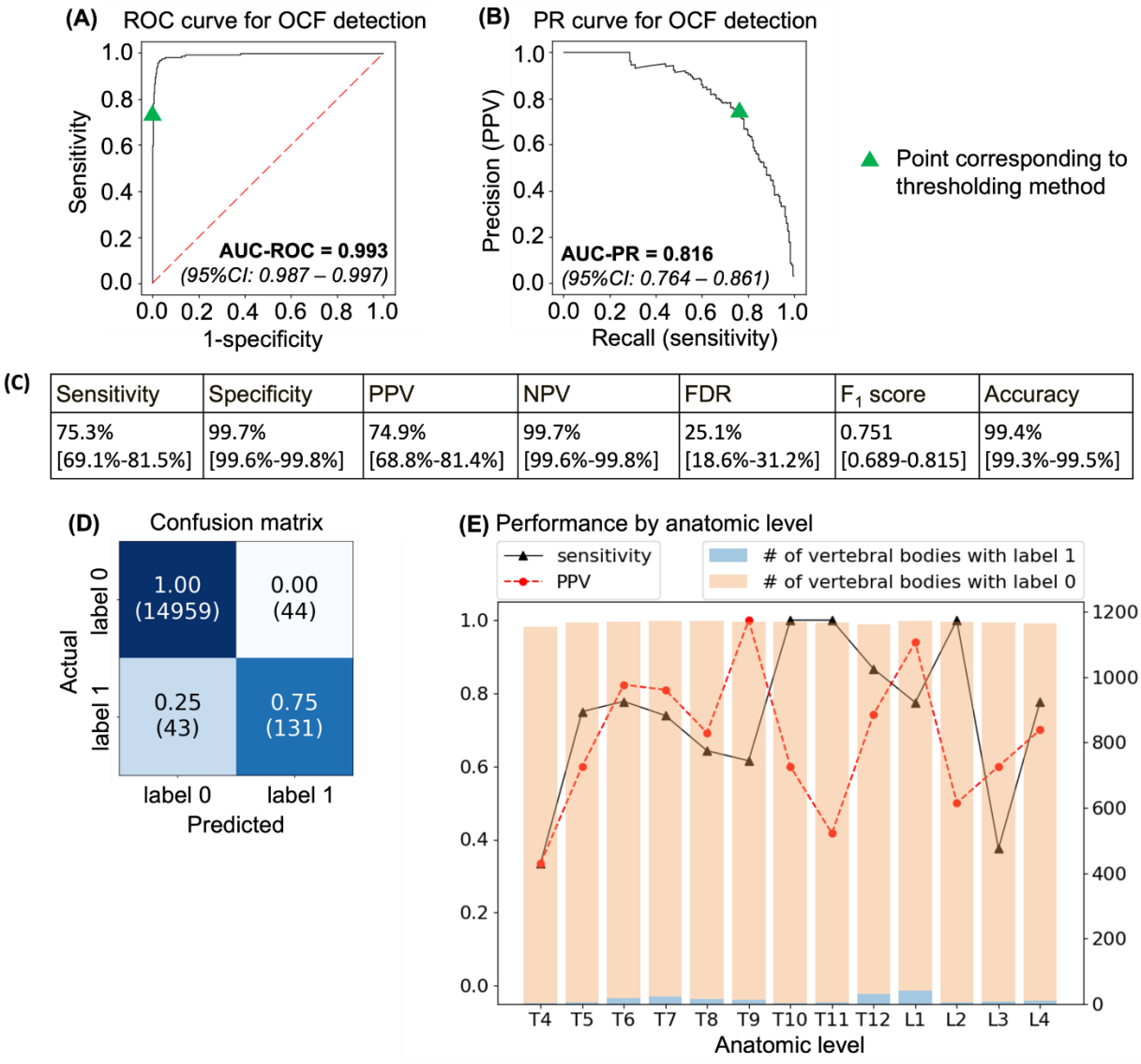


(C)

Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
68.4% [63.3%-73.4%]	98.6% [98.3%-98.9%]	69.5% [64.2%-74.8%]	98.5% [98.2%-98.8%]	30.5% [25.2%-35.8%]	0.689 [0.637-0.741]	97.2% [96.9%-97.6%]



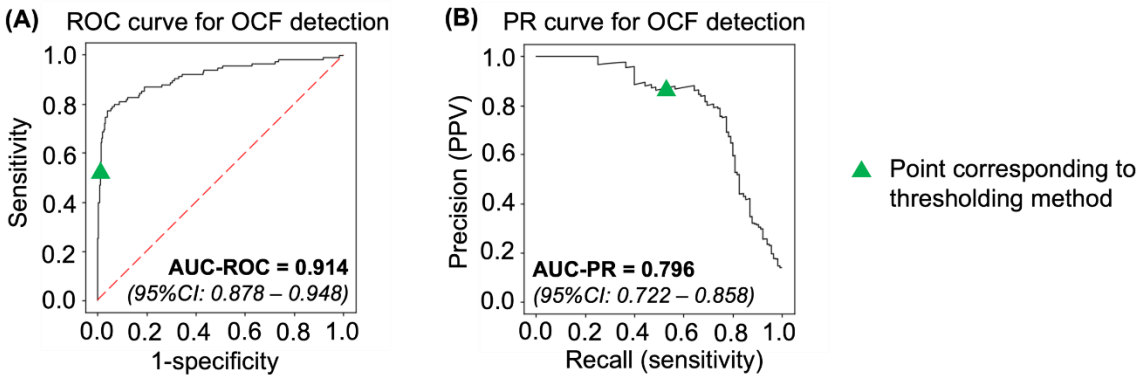
Supplementary Figure 32. Performance of the model built using Inception-ResNet-v2 in Task 3 and evaluated on the test set of the local-m2ABQ dataset.



(C)

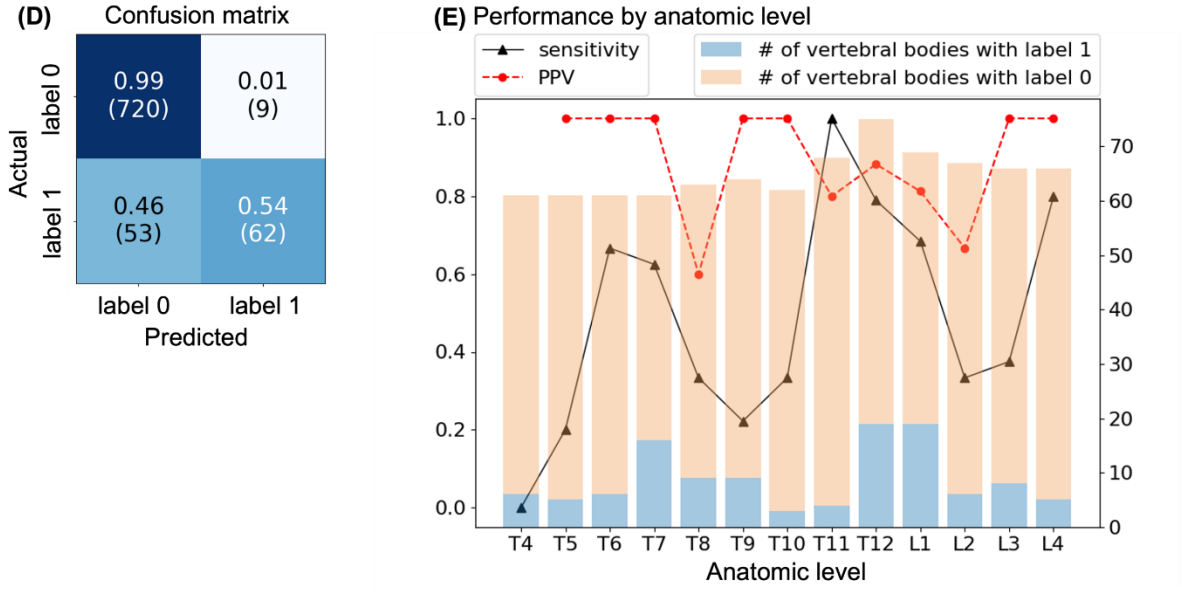
Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
75.3% [69.1%-81.5%]	99.7% [99.6%-99.8%]	74.9% [68.8%-81.4%]	99.7% [99.6%-99.8%]	25.1% [18.6%-31.2%]	0.751 [0.689-0.815]	99.4% [99.3%-99.5%]

Supplementary Figure 33. Performance of the model built using EfficientNet-B1 in Task 1 and evaluated on the test set of the MrOS-mSQ dataset.

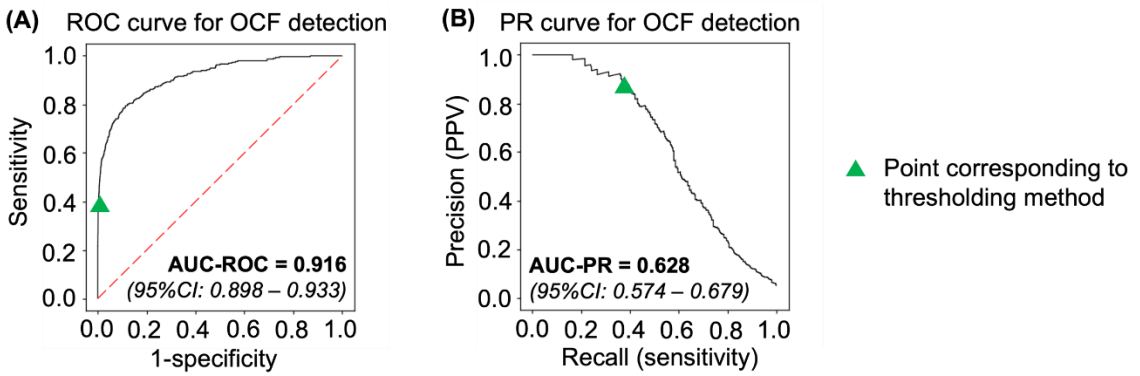


(C)

Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
53.9% [44.2%-63.2%]	98.8% [97.9%-99.5%]	87.3% [79.4%-94.6%]	93.1% [91.4%-94.9%]	12.7% [5.4%-20.6%]	0.667 [0.568-0.757]	92.7% [90.9%-94.3%]

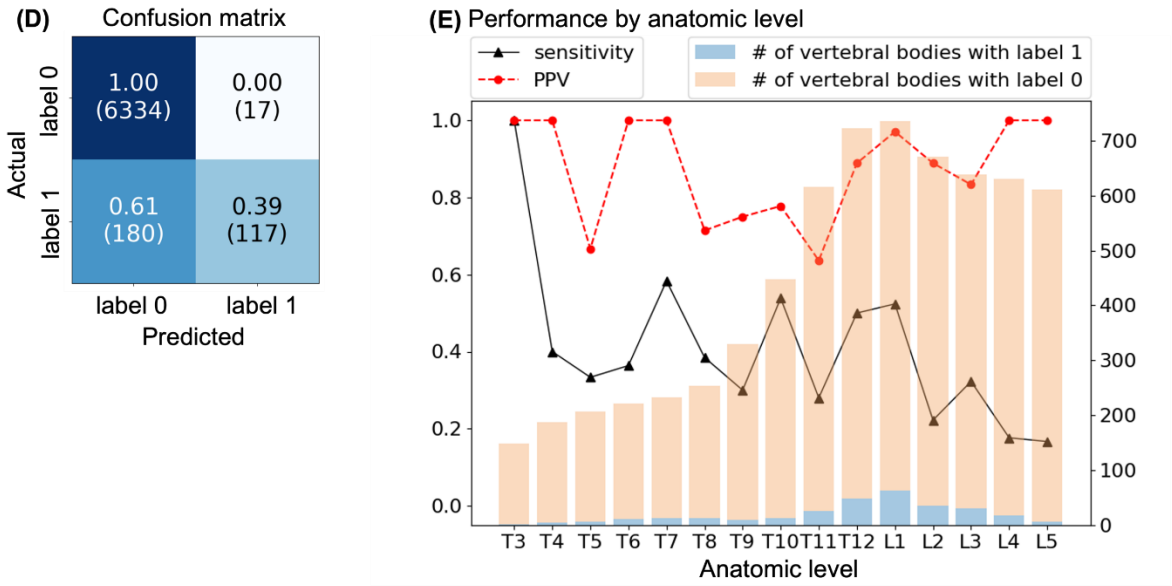


Supplementary Figure 34. Performance of the model built using EfficientNet-B1 in Task 1 and evaluated on the test set of the MrOS-m2ABQ dataset.

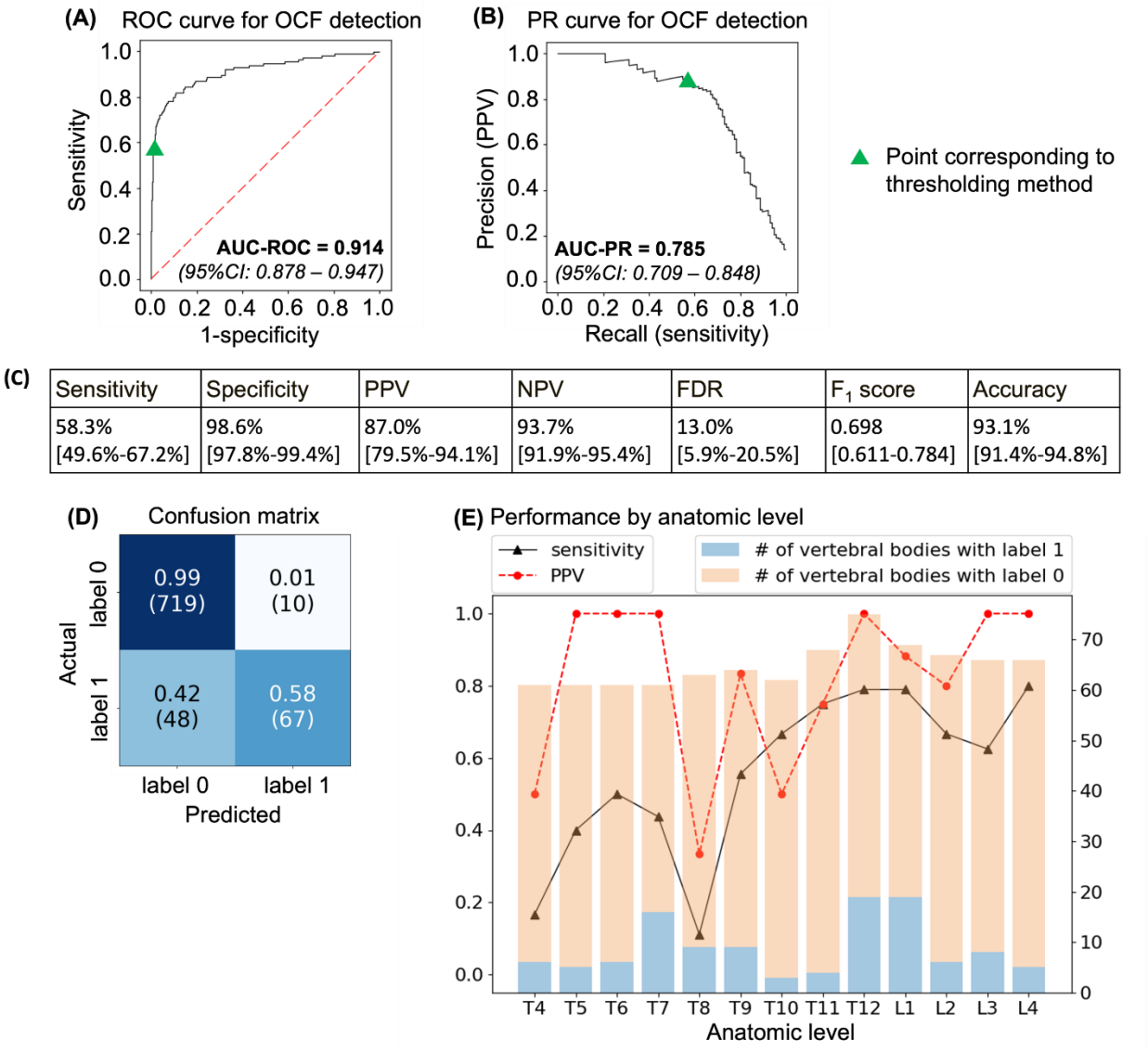


(C)

Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
39.4% [34.0%-45.0%]	99.7% [99.6%-99.8%]	87.3% [81.6%-92.6%]	97.2% [96.8%-97.6%]	12.7% [7.4%-18.4%]	0.543 [0.481-0.606]	97.0% [96.6%-97.4%]

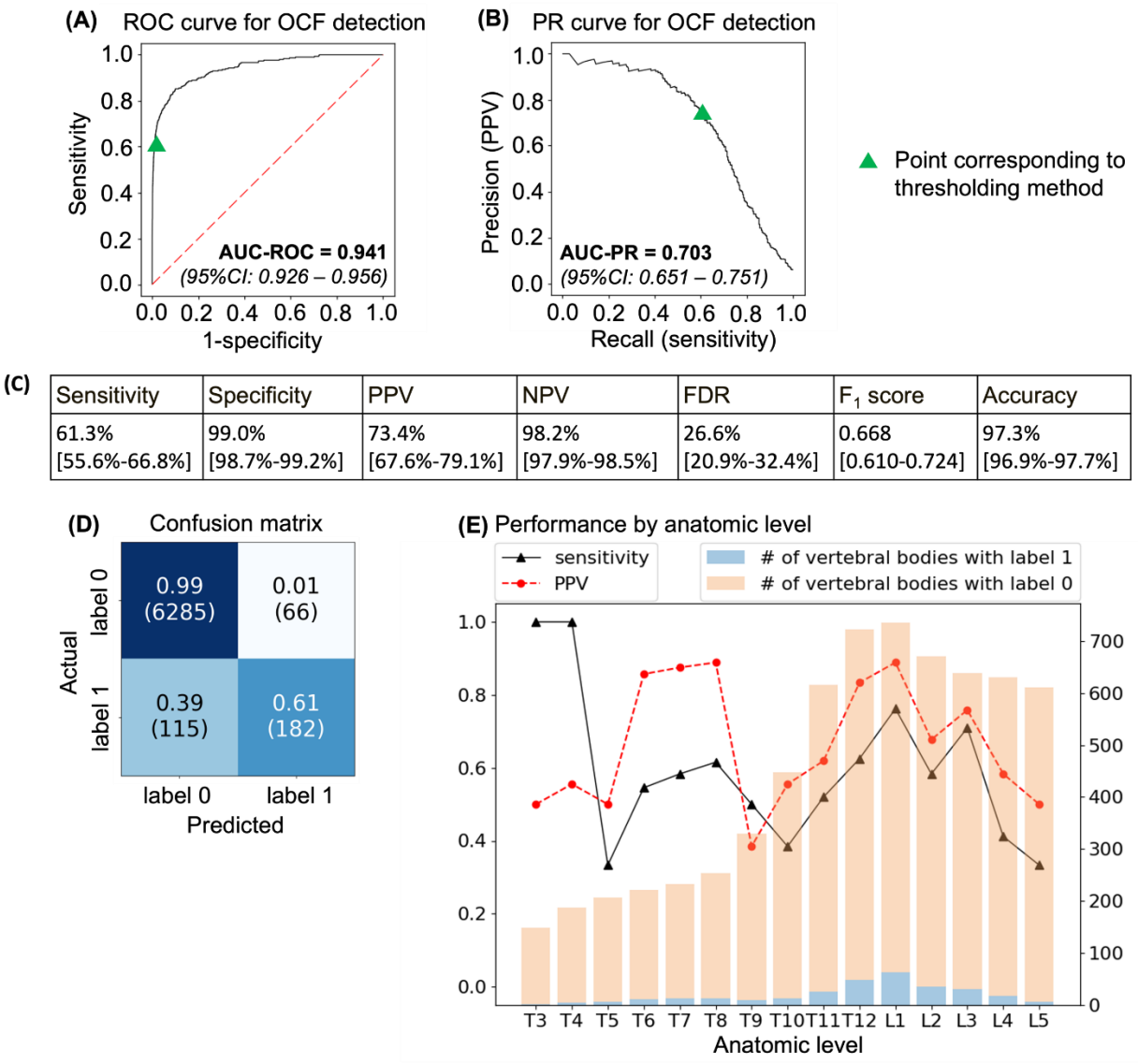


Supplementary Figure 35. Performance of the model built using EfficientNet-B1 in Task 1 and evaluated on the test set of the local-m2ABQ dataset.



Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
58.3% [49.6%-67.2%]	98.6% [97.8%-99.4%]	87.0% [79.5%-94.1%]	93.7% [91.9%-95.4%]	13.0% [5.9%-20.5%]	0.698 [0.611-0.784]	93.1% [91.4%-94.8%]

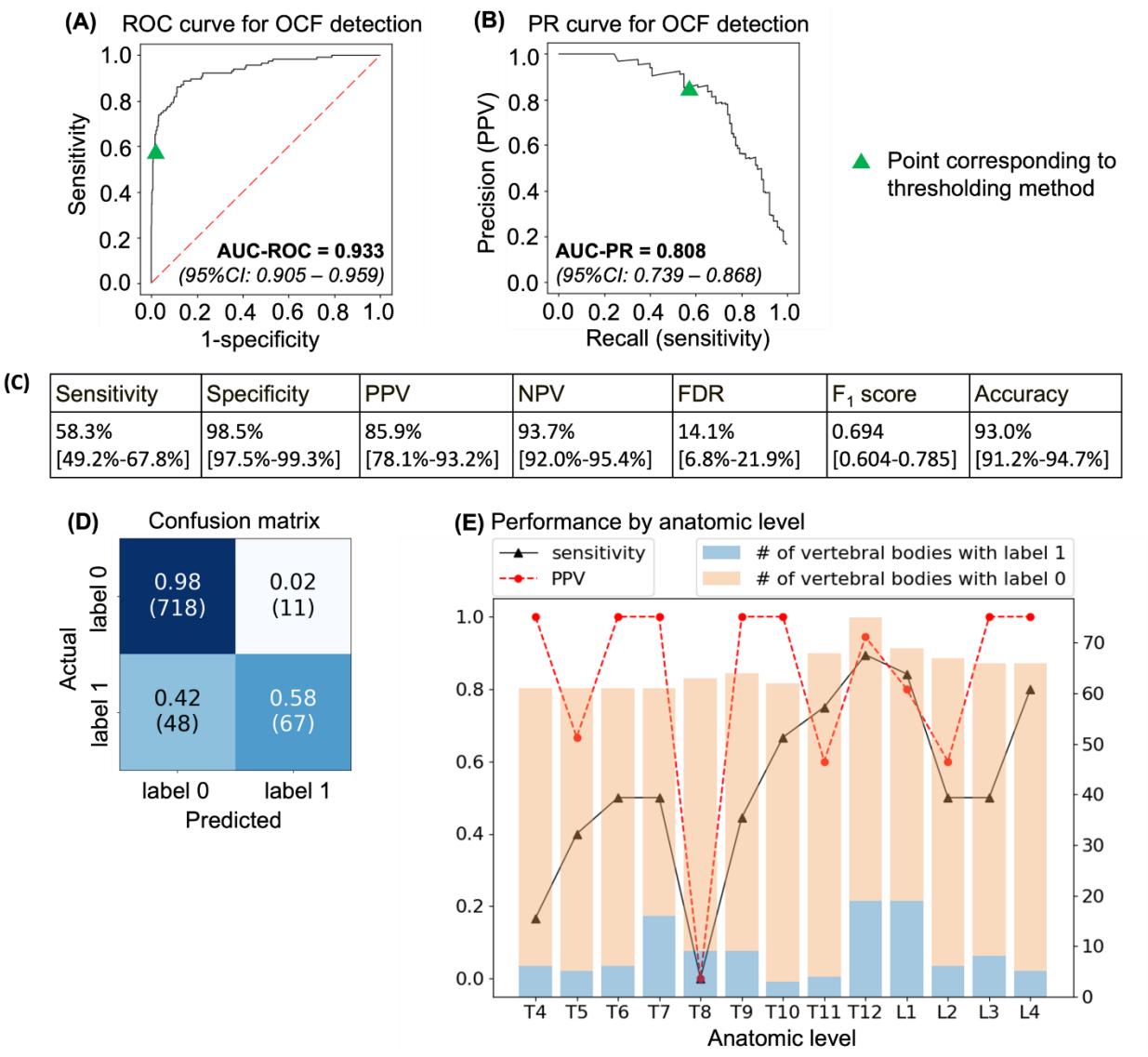
Supplementary Figure 36. Performance of the model built using EfficientNet-B1 in Task 2 and evaluated on the test set of the MrOS-m2ABQ dataset.



(C)

Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
61.3% [55.6%-66.8%]	99.0% [98.7%-99.2%]	73.4% [67.6%-79.1%]	98.2% [97.9%-98.5%]	26.6% [20.9%-32.4%]	0.668 [0.610-0.724]	97.3% [96.9%-97.7%]

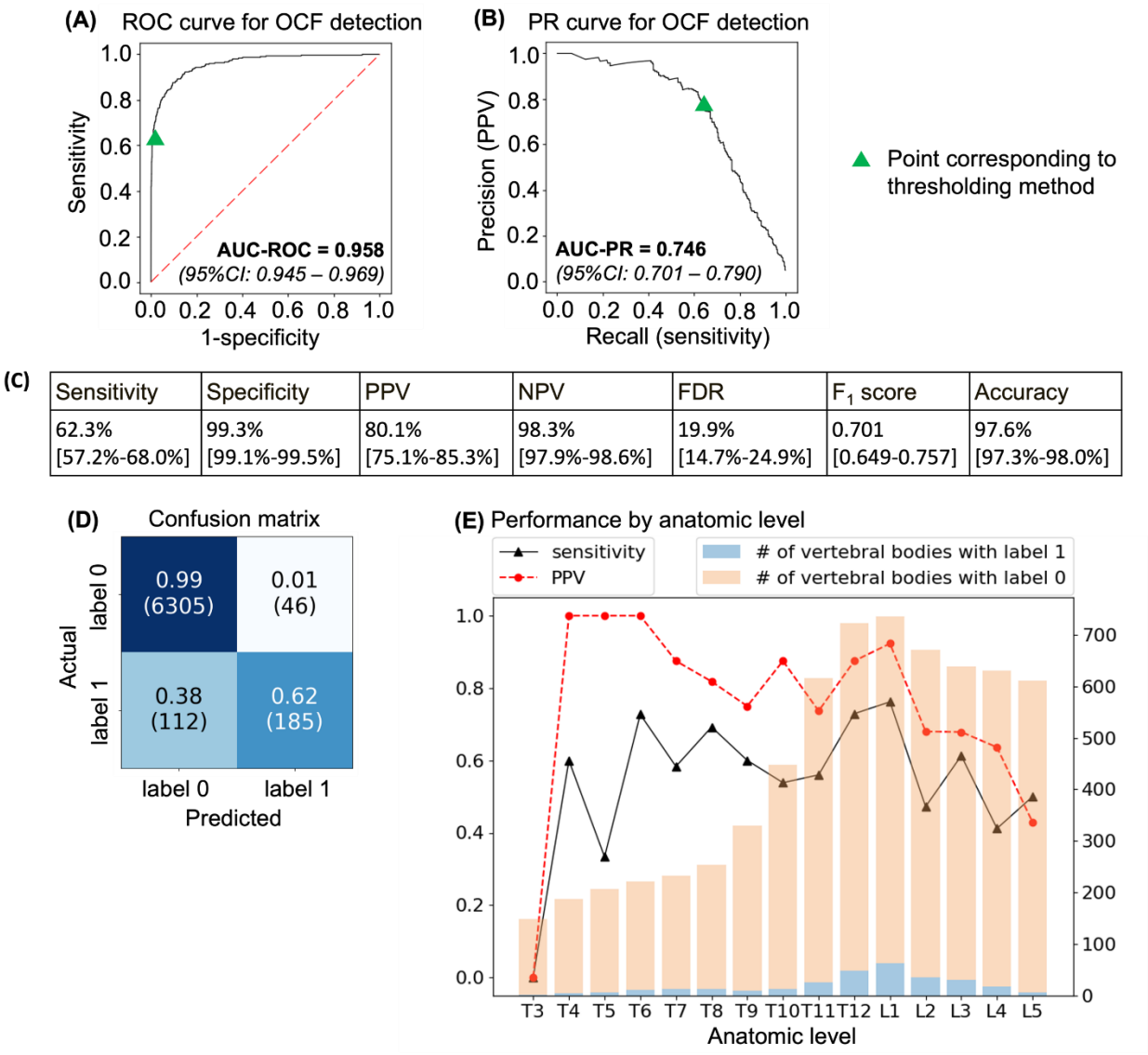
Supplementary Figure 37. Performance of the model built using EfficientNet-B1 in Task 2 and evaluated on the test set of the local-m2ABQ dataset.



(C)

Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
58.3% [49.2%-67.8%]	98.5% [97.5%-99.3%]	85.9% [78.1%-93.2%]	93.7% [92.0%-95.4%]	14.1% [6.8%-21.9%]	0.694 [0.604-0.785]	93.0% [91.2%-94.7%]

Supplementary Figure 38. Performance of the model built using EfficientNet-B1 in Task 3 and evaluated on the test set of the MrOS-m2ABQ dataset.



(C)

Sensitivity	Specificity	PPV	NPV	FDR	F_1 score	Accuracy
62.3% [57.2%-68.0%]	99.3% [99.1%-99.5%]	80.1% [75.1%-85.3%]	98.3% [97.9%-98.6%]	19.9% [14.7%-24.9%]	0.701 [0.649-0.757]	97.6% [97.3%-98.0%]

Supplementary Figure 39. Performance of the model built using EfficientNet-B1 in Task 3 and evaluated on the test set of the local-m2ABQ dataset.

D. Number and percentage of radiographs generated by each type of machine

Supplementary Table 5 shows the number and the percentage of radiographs generated by each type of machine in each of the training, validation, and test sets of the local dataset, as well as the entire local dataset.

Supplementary Table 5. Number and percentage of radiographs generated by each type of machine in each of the training, validation, and test sets of the local dataset, as well as the entire local dataset.

	Training set	Validation set	Test set	Entire local dataset
	Number (percentage) of radiographs			
Canon				
CXDI	5 (0.7%)	0 (0%)	5 (0.6%)	10 (0.5%)
DeJarnette Research Systems				
ImageShare CR	48 (6.5%)	21 (6.8%)	48 (5.7%)	117 (6.2%)
Fujifilm				
5000	72 (9.7%)	22 (7.2%)	61 (7.2%)	155 (8.2%)
5000R	0 (0%)	2 (0.7%)	1 (0.1%)	3 (0.2%)
5501D	20 (2.7%)	9 (2.9%)	16 (1.9%)	45 (2.4%)
AC-3CS	1 (0.1%)	1 (0.3%)	4 (0.5%)	6 (0.3%)
Specific machine name not recorded	285 (38.4%)	123 (40.1%)	345 (40.7%)	753 (39.7%)
General Electric				
Discovery XR656	71 (9.5%)	30 (9.8%)	89 (10.5%)	190 (10.0%)
Revolution XQi ADS_28.2	0 (0%)	1 (0.3%)	1 (0.1%)	2 (0.1%)
Revolution XRd ADS_27.5	2 (0.3%)	0 (0%)	1 (0.1%)	3 (0.2%)
Revolution XRd ADS_28.2	14 (1.9%)	1 (0.3%)	18 (2.1%)	33 (1.7%)
Thunder Platform	112 (15.1%)	40 (13.0%)	121 (14.3%)	273 (14.4%)
WDR1	3 (0.4%)	2 (0.7%)	2 (0.2%)	7 (0.4%)
Philips				
Digital Diagnost	104 (14.0%)	50 (16.3%)	127 (15.0%)	281 (14.8%)
Hybrid General Electric and Fujifilm				
Specific machine name not recorded	5 (0.7%)	5 (1.6%)	8 (1.0%)	18 (0.9%)

E. Demographic information of the subject in the training set of the MrOS dataset

We balanced the training set of the MrOS dataset by downsampling the data instances with label 0 (35). Supplementary Table 6 shows the demographic information of the subject in the training set of the MrOS dataset before and after this downsampling step. This demographic information is also shown in our previous work (35).

Supplementary Table 6. Demographic information of the subject in the training set of the MrOS dataset before and after downsampling the data instances with label 0.

	Training set before downsampling the data instances with label 0	Training set after downsampling the data instances with label 0

	Mean ± standard deviation	
Age at Visit 1	73.7 ± 5.9	73.7 ± 5.8
Age at Visit 2	77.8 ± 5.6	77.9 ± 5.6
	Number (percentage) of subjects	
Race/ethnicity		
American Indian or Alaska Native	42 (0.8%)	14 (0.7%)
Asian	159 (3.2%)	46 (2.5%)
Black or African American	212 (4.2%)	56 (3.0%)
Hispanic or Latino	100 (2.0%)	38 (2.0%)
Native Hawaiian or Other Pacific Islander	11 (0.2%)	5 (0.3%)
White	4,492 (89.6%)	1,715 (91.5%)
	Number	
Total recorded races/ethnicities	5,016	1,874

F. Supplemental References

51. Gonzalez RC, Woods RE. Digital Image Processing. 4th ed. New York, NY: Pearson 2018.

52. Buades A, Coll B, Morel JM. Non-local means denoising. Image Processing On Line. 201;1:208-212.

53. Image denoising. OpenCV. https://docs.opencv.org/3.4/d5/d69/tutorial_py_non_local_means.html. Accessed April 2, 2022.

54. Normalization (image processing). Wikipedia. https://en.wikipedia.org/wiki/Normalization_(image_processing). Accessed April 2, 2022.

55. erode(). OpenCV. https://docs.opencv.org/3.4/d4/d86/group__imgproc__filter.html#gaeb1e0c1033e3f6b891a25d0511362aeb. Accessed April 12, 2022.

56. TensorFlow. https://www.tensorflow.org. Accessed April 12, 2022.

57. Silberman N, Guadarrama S. TF-Slim: a high level library to define complex models in TensorFlow. Google AI Blog. https://ai.googleblog.com/2016/08/tf-slim-high-level-library-to-define.html. Published August 30, 2016. Accessed April 12, 2022.

58. Khan S, Rahmani H, Shah SA, et al. A Guide to Convolutional Neural Networks for Computer Vision. Morgan & Claypool 2018.
59. TensorFlow Model Garden. GitHub. https://github.com/tensorflow/models. Updated July 24, 2020. Accessed April 12, 2022.
60. EfficientNet. GitHub. https://github.com/qubvel/efficientnet. Accessed April 12, 2022.
61. He K, Zhang X, Ren S, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proceedings of ICCV, Santiago, Chile. Washington, D.C.: IEEE Computer Society, 2015; 1026-1034.
62. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: Proceedings of ICLR, San Diego, CA. New York, NY: Association for Computing Machinery, 2015.
63. tf.nn.weighted_cross_entropy_with_logits. TensorFlow. https://www.tensorflow.org/api_docs/python/tf/nn/weighted_cross_entropy_with_logits. Accessed April 12, 2022.
64. Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge, MA: MIT press 2016.
65. inception_v1.py. GitHub. https://github.com/tensorflow/models/blob/master/research/slim/nets/inception_v1.py. Accessed April 12, 2022.
66. inception_resnet_v2.py. GitHub. https://github.com/tensorflow/models/blob/master/research/slim/nets/inception_resnet_v2.py. Accessed April 12, 2022.