Development and validation of a deep learning algorithm detecting 10 common abnormalities on chest radiographs

Ju Gang Nam1,2, Minchul Kim3, Jongchan Park3, Eui Jin Hwang1,2, Jong Hyuk Lee1,2, Jung Hee Hong1,2, Jin Mo Goo1,2,4 and Chang Min Park1,2,4

Affiliations: 1Dept of Radiology, Seoul National University Hospital, Seoul, Republic of Korea. 2College of Medicine, Seoul National University, Seoul, Republic of Korea. 3Lunit Inc., Seoul, Republic of Korea. 4Institute of Radiation Medicine, Seoul National University Medical Research Center, Seoul, Republic of Korea.

Correspondence: Chang Min Park, Dept of Radiology and Institute of Radiation Medicine, Seoul National University College of Medicine, 101 Daehak-ro, Jongno-gu, Seoul 03080, Republic of Korea.
E-mail: cmpark.morphius@gmail.com

ABSTRACT

We aimed to develop a deep learning algorithm detecting 10 common abnormalities (DLAD-10) on chest radiographs, and to evaluate its impact in diagnostic accuracy, timeliness of reporting and workflow efficacy.

DLAD-10 was trained with 146 717 radiographs from 108 053 patients using a ResNet34-based neural network with lesion-specific channels for 10 common radiological abnormalities (pneumothorax, mediastinal widening, pneumoperitoneum, nodule/mass, consolidation, pleural effusion, linear atelectasis, fibrosis, calcification and cardiomegaly). For external validation, the performance of DLAD-10 on a same-day computed tomography (CT)-confirmed dataset (normal:abnormal 53:147) and an open-source dataset (PadChest; normal:abnormal 339:334) was compared with that of three radiologists. Separate simulated reading tests were conducted on another dataset adjusted to real-world disease prevalence in the emergency department, consisting of four critical, 52 urgent and 146 nonurgent cases. Six radiologists participated in the simulated reading sessions with and without DLAD-10.

DLAD-10 exhibited area under the receiver operating characteristic curve values of 0.895–1.00 in the CT-confirmed dataset and 0.913–0.997 in the PadChest dataset. DLAD-10 correctly classified significantly more critical abnormalities (95.0% (57/60)) than pooled radiologists (84.4% (152/180); p=0.01). In simulated reading tests for emergency department patients, pooled readers detected significantly more critical (70.8% (17/24) versus 29.2% (7/24); p=0.006) and urgent (82.7% (258/312) versus 78.2% (244/312); p=0.04) abnormalities when aided by DLAD-10. DLAD-10 assistance shortened the mean±SD time-to-report critical and urgent radiographs (640.5±466.3 versus 3371.0±1352.5 s and 1840.3±1141.1 versus 2127.1±1468.2 s, respectively; all p<0.01) and reduced the mean±SD interpretation time (20.5±22.8 versus 23.5±23.7 s; p<0.001).

DLAD-10 showed excellent performance, improving radiologists’ performance and shortening the reporting time for critical and urgent cases.