subarticular recesses, or intervertebral foramen, which in turn causes compression of the associated neural structures [6].

While LSS is defined as a clinical syndrome and diagnosed as such, imaging is frequently used as a confirmatory tool prior to determining treatment. Magnetic resonance imaging (MRI) is the mainstay modality for detailed anatomical assessment of the spine with excellent soft tissue contrast and used for confirmation of clinical symptom-based diagnoses and determining the optimal treatment course. MRI is essential for evaluation of LSS and to confirm clinical findings, largely supplanting CT myelogram studies due to its superior soft tissue contrast. A number of studies have attempted to identify core quantitative radiological criteria for the diagnosis of LSS [7–9]. However, lumbar spine MRI interpretation is time-intensive and depends on the individual radiologist or surgeons' expertise and experience, therefore supporting objective and standardized methods of diagnosing and decision-making are desired.

Machine learning (ML) models, including deep convolutional neural networks (CNN), have already been successfully applied for evaluation of LSS and other degenerative changes with high accuracy in various approaches [10, 11]. However, most CNN algorithms rely on one-component models for binary classification (present/absent) of LSS. One recent work applied a two-component CNN to detect stenosis targeting the central canal, lateral recesses, and neural foramina with subsequent grading of the stenosis [12].

Here, the purpose of this study was to develop a three-stage convolutional neural network (CNN) approach to segment anatomical structures, classify the presence of lumbar spinal stenosis (LSS) and assess its severity on spine MRI on axial and sagittal MR images. The classification covers the detection of all three stenosis types—central canal, foraminal, and lateral recess. The performance of the model has been compared to a panel of radiologist subspecialists to test its reliability and accuracy.

Methods

Data set and annotation

External institutional review board approval was obtained to retrospectively review anonymized imaging data. The initial data set consisted of 1635 MRI studies of adult subjects referred for lumbar spine MRI for low back pain. The data set consisted of 45.7% of males (54.3% of females), with age ranging from 18 to 85 years. Each MRI study corresponded to a patient (1635 MRI studies = 1635 patients). Patients with implants or instrumentation, severe scoliosis, and poor image quality were excluded. Each MRI study was acquired using a standard lumbar spine protocol, including T2-weighted axial and sagittal pulse sequences with balanced labels. First, T2-weighted axial sequence was extracted from each lumbar study. All slices from the lumbar disc levels (L1/L2, L2/L3, L3/L4, L4/L5, L5/S1) were selected from the T2-weighted axial sequence and labeled per slice. On average, around 10 to 15 axial slices were obtained from each study, with a more precise count of approximately 13.3 for this specific dataset (21,702 images in total).

For axial images, muscle tissue, the discs, spinal canal, thecal sac, neural foramina, nerves, nerve roots, lateral recess, facet joints, spinous process, articular process, ligamentum flavum, disc bulging or herniation, arteries, veins, and kidneys were labeled. On sagittal plane images, the discs, vertebral body, spinal canal, spinal cord with nerve roots, and spinous processes were labeled for segmentation. The segmentation labelling was performed by administrators.

In addition to segmentation of the key anatomical structures, the studies were also labeled by musculoskeletal-trained radiologist subspecialists on a scale of 0 (absent), 1 (mild), 2 (moderate), 3 (severe) for LSS to establish the reference standard [13]. The reference standard was determined by majority voting rule and in case of disagreement, adjudicated by a further radiologist. At first, during training, the segmentation model detected the facet joints and spinal canal. Then images were resized and augmented (horizontal flips and rotations). Of the 1635 studies, 1390 were used for CNN weight training, and 245 as a validation set for hyperparameter tuning. The dataset was randomly divided into a training set for training the CNN parameters (1635 studies) and a validation set for hyperparameter tuning (245 studies). This was done while maintaining the same distribution in each subset, split ratio and the rule that images from one study can't be in different sets at the same time. Thus, data splitting was carried out based on studies rather than individual images, such as images from the same MRI study could go to different subsets (to avoid for instance two images from the same MRI study going to both training and testing).

Additionally, an external data set of 150 studies were reserved for final model accuracy assessment. The inclusion/exclusion criteria were identical to the those of the training data set. The external validation set were graded on a scale of absent, mild, moderate or severe by a panel of 7 radiologist subspecialists. The radiologist interpretations were then compared to the interpretation of the model.

The number of images for each stenosis (central, lateral recess, foraminal) for each severity class (absence, mild, moderate, severe) and for each data subset (train, validation, test) is shown in Supplementary Table 1.