Nakamoto *et al.*
BioMedical Engineering OnLine (2025) 24:11
https://doi.org/10.1186/s12938-025-01341-4

BioMedical Engineering OnLine

RESEARCH Open Access

WDRIV-Net: a weighted ensemble transfer learning to improve automatic type stratification of lumbar intervertebral disc bulge, prolapse, and herniation

Ichiro Nakamoto[1†], Hua Chen[4,5†], Rui Wang[2,3], Yan Guo[1], Wei Chen[1], Jie Feng[4,5] and Jianfeng Wu[2,3*]

†Ichiro Nakamoto and Hua Chen contributed equally to the research.

*Correspondence: wjf1801@qq.com

[1] School of Internet Economics and Business, Fujian University of Technology, Fuzhou, China
[2] Department of Neurosurgery, Pingtan Comprehensive Experimentation Area Hospital, Pingtan, China
[3] Department of Neurosurgery, Fujian Medical University Union Hospital, Fuzhou, China
[4] Department of Radiology, Pingtan Comprehensive Experimentation Area Hospital, Pingtan, China
[5] Department of Radiology, Fujian Medical University Union Hospital, Fuzhou, China

Abstract

The degeneration of the intervertebral discs in the lumbar spine is the common cause of neurological and physical dysfunctions and chronic disability of patients, which can be stratified into single—(e.g., disc herniation, prolapse, or bulge) and comorbidity-type degeneration (e.g., simultaneous presence of two or more conditions), respectively. A sample of lumbar magnetic resonance imaging (MRI) images from multiple clinical hospitals in China was collected and used in the proposal assessment. We devised a weighted transfer learning framework WDRIV-Net by ensembling four pre-trained models including Densenet169, ResNet101, InceptionV3, and VGG19. The proposed approach was applied to the clinical data and achieved 96.25% accuracy, surpassing the benchmark ResNet101 (87.5%), DenseNet169 (82.5%), VGG19 (88.75%), InceptionV3 (93.75%), and other state-of-the-art (SOTA) ensemble deep learning models. Furthermore, improved performance was observed as well for the metric of the area under the curve (AUC), producing a ≥ 7% increase versus other SOTA ensemble learning, a ≥ 6% increase versus most-studied models, and a ≥ 2% increase versus the baselines. WDRIV-Net can serve as a guide in the initial and efficient type screening of complex degeneration of lumbar intervertebral discs (LID) and assist in the early-stage selection of clinically differentiated treatment options.

Keywords: Lumbar spine degeneration, Lumbar intervertebral discs, Transfer learning, Deep learning, Magnetic resonance imaging, Weighted ensemble learning

Introduction

Lumbar intervertebral discs (LID) are essential for spinal motion, flexibility, and stability [1]. Lumbar disc herniation (LDH), lumbar disc prolapse (LDP), and lumbar disc bulge (LDB) are three prevailing types of lumbar spine degeneration observed in clinical practice worldwide [1–5]. Degenerative changes in the LID including LDH, LDP, and LDB can occur as early as the first decade of life and can be associated with symptoms including severe/chronic lower back pain, sciatica, muscle spasms, and disability impacting the quality of life [1–3]. For patients simultaneously experiencing

multiple conditions (i.e., comorbidity), the disease may advance to a more severe or complicated stage, causing generally a situation where differentiated health care or urgent intervention/treatment is needed [6, 7]. These patients are more likely to bear more stress and costs if the decision-making is sub-optimal. Clinically, it is of importance to automatically identify the more complicated cases (e.g., comorbidity-type) from those less complicated counterparts (e.g., single-type) at an earlier stage effectively.

Radiography images using X-, gamma-ray techniques, computerized tomography scans (CT), and magnetic resonance imaging (MRI) are primarily exploited in detecting the disease of LID [8, 9]. However, MRI is the preferred imaging modality for LID-related diseases versus other imaging techniques [9] and is widely utilized in treatment planning [8–11]. Accurate interpretation of the principal traits of MRI of LID is crucial for the preprocedural assessment of potential interventions for the lumbar spine such as rehabilitation training, injection-based treatment, and surgery [12]. Presently manual inspection by medical professionals represents the most commonplace approach to extracting information from MRI images [8]. The visual examination carried out slide-by-slide by experts typically relies on the expertise of medical professionals and mostly is time-consuming and bias-prone [6, 12]. Challenges characterized by ambiguity, inconsistency, or conflict of diagnosis are not uncommon, which may induce unintended medical consequences [6]. Therefore, the challenge of automating the classification of disc degeneration remains a significant concern for both patients and physicians [6, 12].

The diagnosis of LID degeneration is also influenced by factors such as image quality and analysis techniques, beyond the inherent bias of manual observation [10, 11, 13]. These variables significantly impact the evaluation of LID degeneration type and progression, as well as subsequent treatment decisions [5, 11, 14]. Specifically, as shown in Fig. 1, aeolotropic dimensional resolution, inter-type similarity, and intra-type variety of MRI of LID may yield difficulty in identifying the type or severity of degeneration [15–17]. Therefore, designing an automatic classification framework for LID degeneration is meaningful to address the issue of sub-optimal image quality, enhance the efficiency of physician diagnosis, optimize the operability level of LID clinical applications, and alleviate the financial, physical as well as psychiatric burden on patients [10, 11, 18, 19].

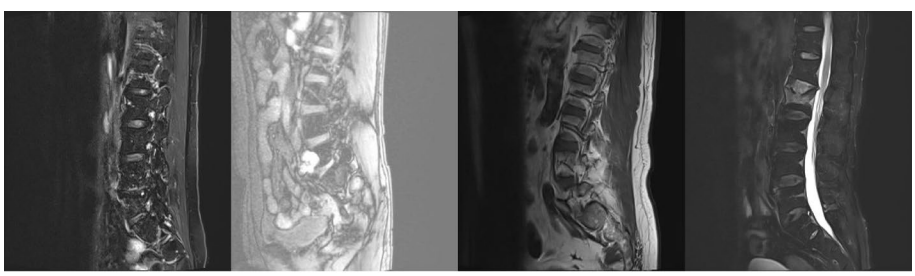

Fig. 1 Aeolotropic dimensional resolution, inter-type similarity, and intra-type variety of lumbar spine MRI (**a**–**d**) induce challenges in detecting the degeneration type of LID including bulge, prolapse, and herniation

The major contributions of this research are outlined as follows:

1. The proposed framework employs a weighted ensemble of fine-tuned benchmark models using a deep transfer learning approach to improve performance and generalizability in medical image processing.
2. The proposed approach aims to precisely identify and stratify single- and comorbidity-type degeneration in the MRI images of LID at an early phase that has not been investigated in previous literature, addressing the problems in traditional methodologies such as false-positive rates, false-negative rates, data noise (e.g., aeolotropic dimensional resolution, inter-type similarity, and intra-type variety), accuracy, and area under the curve value (AUC) of a sub-optimal single classifier.
3. The proposed ensemble model employs pre-processing steps, including data augmentation and denoising, to counter the challenges of resolution variations and data noises, which potentially bias the accuracy of stratification. The knowledge learned from ImageNet is effectively transferred and guides the model to the correct targets by yielding improved outputs.
4. The effectiveness and robustness of the proposed approach are evaluated using a variety of metrics including AUC, accuracy, recall, precision, F1-score, and confusion matrix. Further, the performance of the proposed framework is compared with and superior to the outcomes of the benchmarks, recent most-studied models, and other SOTA deep learning.

The remaining sections of the study are organized as follows: Section Literature review delineates literature work investigating machine learning, deep learning, and ensemble learning with medical image tasks and convolution-network processing. Section Results depicts the experimental tests. Section Discussion and conclusions concludes the research with research limitations and future directions. And section Materials and methods explains the proposed methodology.

Literature review

Machine learning and deep neural networks have proven effective in addressing the complexity and variability of medical image analysis [20–22]. In the last decade, numerous semi- or full-supervised algorithms have been devised to analyze medical images and assist medical diagnosis driven by the advancement of imaging quality and analytical techniques [18, 19, 21, 23–33]. Full supervised learning generally uses labeled data, whereas the semi-supervised counterpart normally regulates the output of deep learning.

Ensemble learning and deep learning are the predominant machine learning approaches that have witnessed applications in medical image-related tasks during recent years, achieving benchmark performance across various disease-related tasks [16–19, 28, 32, 34–40]. Deep learning has made substantial advancements in image analysis during the past decade, excelling at feature extraction [16, 17, 21, 25–27]. These techniques, particularly when combined with convolutional neural networks (CNNs), have been instrumental in assisting medical professionals with diagnosis.

CNNs, evolving over the past half a century, are crucial in image classification [37–40], object detection [15, 21, 41], and segmentation [34, 42–46].

On the other hand, challenges and prospects in medical image processing have been widely discussed [16–19, 47–53]. For example, Tavana et al. [16] demonstrated the potential of deep learning in classifying the type of spinal curvature. Pandi et al. [47] and Niu et al. [48] highlighted the use of evolutionary deep full CNNs in medical image segmentation. Tanveer et al. [20] exploited deep learning to analyze speech signal tasks. In addition, Niu et al. [48] explored application scenarios other than medical image processing using deep learning. Additionally, Zheng et al. [53] discussed the use of deep learning techniques to detect image-level classification for breast histopathology, and XGBoost feature selection has been used to improve protein–protein interaction prediction accuracy [18, 19]. Nevertheless, debates remain regarding the challenge of generalizability [16–19, 47–53]. To date, to our best knowledge, no research has addressed the complicated comorbidity degeneration of the lumbar spine in connection with automatic stratification using clinical observations. This underscores the need for a reliable classification algorithm. Developing a robust detection methodology could improve processing efficiency and reduce error rates, motivating our effort to design an approach for automatic stratification of LID degeneration using MRI images.

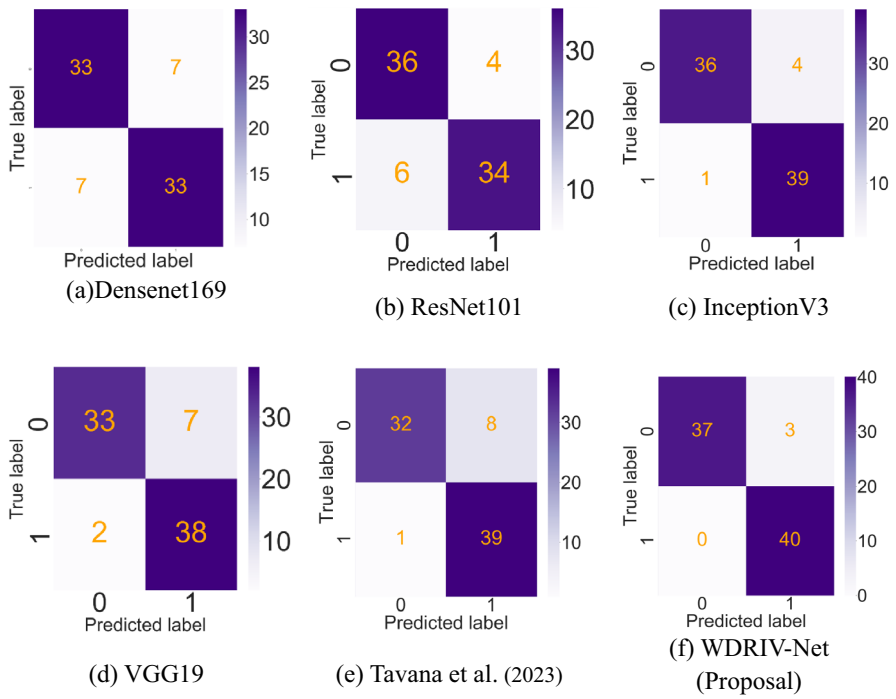

Fig. 2 Confusion matrix for **a** pre-trained Densenet169, **b** pre-trained ResNet101, **c** pre-trained InceptionV3, **d** pre-trained VGG19, **e** Tavana et al. [16, 17], and f the proposed weighted ensemble deep transfer learning model WDRIV-Net. The axis label of 0 represents single-type degeneration of LID, and 1 represents comorbidity-type degeneration of LID

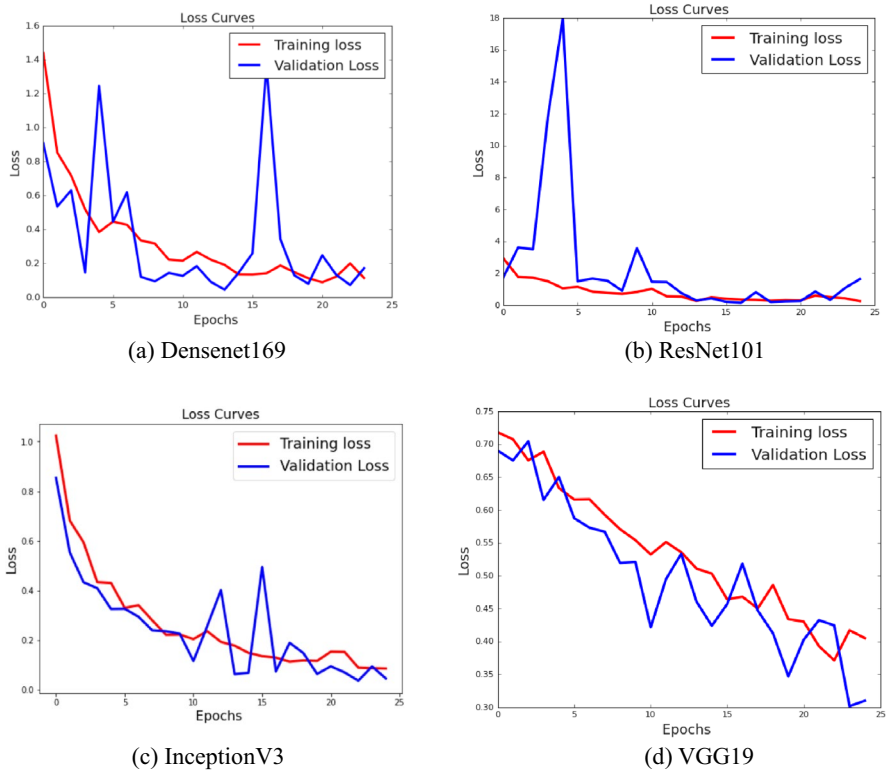

Fig. 3 Trend observation of training loss and validation loss during training for the four baseline models. **a** Densenet169, **b** ResNet101, **c** InceptionV3, **d** VGG19

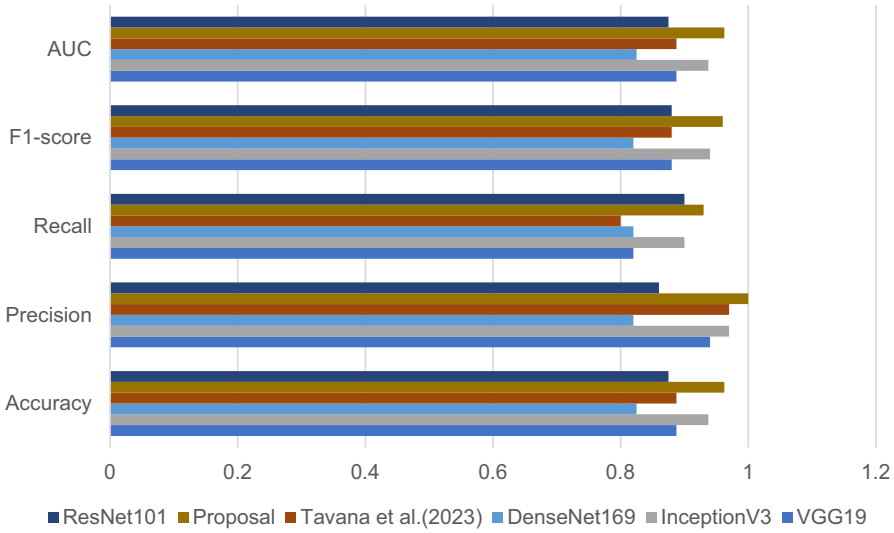

Fig. 4 Evaluation of accuracy, precision, recall, F1-score, and AUC in detecting the single-type degeneration of LID

Results

The comparison of the confusion matrix output for each model, the proposal approach, and other SOTA ensemble learning is shown in Fig. 2. The matrix reveals

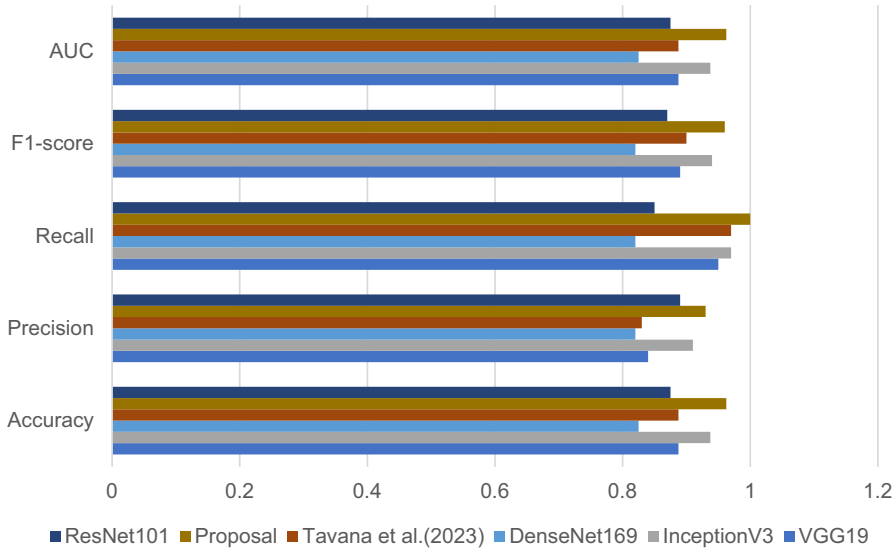

Fig. 5 Evaluation of accuracy, precision, recall, F1-score, and AUC in detecting the comorbidity-type degeneration of LID

true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), verifying the effectiveness and accuracy of the approaches. Training and validation loss trends during training are illustrated in Fig. 3, while accuracy, precision, recall, F1 scores, and AUC of all frameworks, including the proposed WDRIV-Net, are detailed in Figs. 4 and 5.

The confusion matrix suggested that the proposed WDRIV-Net outperformed individual models and other SOTA ensemble frameworks, yielding an accuracy of 96.25% in total (77/80). This surpasses DenseNet169 (82.5%, 66/80), ResNet101 (87.5%, 70/80), VGG19 (88.75%, 71/80), InceptionV3 (93.75%, 75/80), and Tavana et al. (88.75%, 71/80) [16]. Notably, three MRI images of single-type degeneration were misclassified as comorbidity-type degeneration by WDRIV-Net, while no comorbidity-type images were misclassified. Other models, including DenseNet169, ResNet101, VGG19, InceptionV3, and the SOTA model [16], exhibited higher FP and FN rates and lower TP and TN rates compared to WDRIV-Net.

The results in Fig. 3 outlined the trend change in training and validation loss over epochs during training. Although partial of the baseline models (e.g., Densenet169 and ResNet101) observed greater fluctuations, the overall loss trends for all four models were consistent, decreasing with the increment of epochs and approaching zero by the end of training. The optimal validation loss varied among models: ResNet101 observed ~ 0.14 at epoch 17, DenseNet169 ~ 0.05 at epoch 24, InceptionV3 ~ 0.04 at epoch 23, and VGG19 ~ 0.30 at epoch 24.

Figures 4 and 5 delineate the accuracy, precision, recall, F1-score, and AUC for classifying LID degeneration using the proposed WDRIV-Net, benchmark models, and other ensemble methods. The results suggested that WDRIV-Net outperformed all models for both single- and comorbidity-type degeneration, achieving superior accuracy, F1-score, AUC, precision, and recall. Accuracy and AUC followed similar trends

across models, with WDRIV-Net observing the highest values, outperforming pre-trained InceptionV3, ResNet101, DenseNet169, VGG16, and the SOTA model [16]. The weighted ensemble approach improved the performance of individual models and surpassed all single methods and other ensemble strategies. Notably, the ensemble mechanism in [16], which used an equal-weight average, performed worse than the non-equal-weight WDRIV-Net and some baseline models (e.g., InceptionV3).

Further, to evaluate the performance of the proposal WDRIV-Net versus other recent models and ensemble learning, we summarized the outcomes in Table 1 with references to NASNetMobile [52], MobileNetV2 [54], VGG19 [55], NasNetLarge [52], ResNet50 [39], ResNet152 [39], DenseNet121 [42], DenseNet201 [42], the four individual benchmark models integrated into the proposed approach, and the ensemble model by [16]. These reference models represented widely studied or high-performing

Table 1 A comparison between the proposed WDRIV-Net with other models

No	Model	Classifier	Accuracy (%)	Type	Precision	Recall	F1	AUC (%)
1	NASNetMobile [52]	Sigmoid	60.00	Single	0.60	0.60	0.60	60.0
		Sigmoid	60.00	Comorbidity	0.60	0.60	0.60	60.0
2	Pre-trained ResNet152V2	Sigmoid	71.25	Single	1.00	0.42	0.60	71.25
		Sigmoid	71.25	Comorbidity	0.63	1.00	0.78	71.25
3	DenseNet201 [42]	Sigmoid	73.75	Single	1.00	0.47	0.64	73.75
		Sigmoid	73.75	Comorbidity	0.66	1.00	0.79	73.75
4	Pre-trained VGG16	Sigmoid	73.75	Single	0.77	0.68	0.72	73.75
		Sigmoid	73.75	Comorbidity	0.71	0.80	0.75	73.75
5	MobileNetV2 [54]	Sigmoid	76.25	Single	0.92	0.57	0.71	76.25
		Sigmoid	76.25	Comorbidity	0.69	0.95	0.80	76.25
6	ResNet50 [39]	Sigmoid	81.25	Single	0.78	0.88	0.82	81.25
		Sigmoid	81.25	Comorbidity	0.86	0.75	0.80	81.25
7	Pre-trained DenseNet169	Sigmoid	82.50	Single	0.82	0.82	0.82	82.50
		Sigmoid	82.50	Comorbidity	0.82	0.82	0.82	82.50
8	NasNetLarge [52]	Sigmoid	82.50	Single	0.93	0.70	0.80	82.50
		Sigmoid	82.50	Comorbidity	0.76	0.95	0.84	82.50
9	InceptionResNetV2 [23]	Sigmoid	85.00	Single	0.83	0.88	0.85	85.00
		Sigmoid	85.00	Comorbidity	0.87	0.82	0.85	85.00
10	Pre-trained ResNet101	Sigmoid	87.5	Single	0.86	0.90	0.88	87.5
		Sigmoid	87.5	Comorbidity	0.89	0.85	0.87	87.5
11	VGG19 [55]	Sigmoid	88.75	Single	0.94	0.82	0.88	88.75
		Sigmoid	88.75	Comorbidity	0.84	0.95	0.89	88.75
12	ResNet152 [39]	Sigmoid	90.00	Single	0.97	0.82	0.89	90.00
		Sigmoid	90.00	Comorbidity	0.85	0.97	0.91	90.00
13	DenseNet121 [42]	Sigmoid	90.00	Single	0.97	0.82	0.89	90.00
		Sigmoid	90.00	Comorbidity	0.85	0.97	0.91	90.00
14	Pre-trained InceptionV3	Sigmoid	93.75	Single	0.97	0.90	0.94	93.75
		Sigmoid	93.75	Comorbidity	0.91	0.97	0.94	93.75
15	Tavana et al. [16]	Sigmoid	88.75	Single	0.97	0.80	0.88	88.75
		Sigmoid	88.75	Comorbidity	0.83	0.97	0.90	88.75
16	WDRIV-Net (proposed)	Sigmoid	96.25	Single	1.00	0.93	0.96	96.25
		Sigmoid	96.25	Comorbidity	0.93	1.00	0.96	96.25

Table 2 Summary of the private LID MRI sagittal-view data set and parameters

(1) Private data set summary				
Degeneration type	**Description**	**Training sample**	**Testing sample**	**Total sample**
Single-type	(a)Herniation; (b) prolapse; (c) bulge	551	40	591
Comorbidity-type	Combination of symptom (a), (b), and (c)	551	40	591

(2) Hyperparameters for the proposed weighted ensemble approach					
Benchmark	**Image size**	**Train epochs**	**Optimizer**	**Activate**	**Learning rate**
Densenet169	224×224×3	25	Adam	Sigmoid	1e−4
ResNet101	224×224×3	25	Adam	Sigmoid	1e−4
InceptionV3	224×224×3	25	Adam	Sigmoid	1e−4
VGG19	224×224×3	25	Adam	Sigmoid	1e−4

(3) Data augmentation	
Parameter	**Value**
Rotation	20
width_shift_range	0.05
height_shift_range	0.05
horizontal_flip	True
Random drop	0.2

frameworks in other diseases and ensemble techniques [16, 19, 28, 31, 34, 35, 37, 39, 40, 42–45, 52, 54, 56–66].

Table 2 indicates that the proposed ensemble model observed 96.25% accuracy, surpassing [16] (88.75%), DenseNet121 [42] and ResNet152 [39] (90%), InceptionResNetV2 [23] (85%), and all individual baseline models (< 94%). NASNetMobile [52] recorded the lowest accuracy. Similar trends were observed for precision, recall, F1-score, and AUC, with WDRIV-Net achieving the highest AUC. The proposed model demonstrated a ~ 7.5% AUC improvement over [16], a ≥ 6% increase over other top-performing models, and a ≥ 2% improvement over its benchmark models.

Discussion

In this research, we proposed WDRIV-Net, a weighted ensemble-based deep transfer learning framework, to classify LID MRI sagittal-view images from clinical settings, observing 96.25% accuracy for single- and comorbidity-type disc degeneration. WDRIV-Net outperformed other models with a ≥ 7% increase in AUC compared to SOTA ensemble learning, a ≥ 6% improvement over widely studied models, and a ≥ 2% improvement over baseline models (DenseNet169, ResNet101, InceptionV3, and VGG19). Similar gains were observed in precision, recall, and F1-score.

The framework leveraged transfer learning on ImageNet and effectively applied it to a private clinical LID dataset, with pre-processing, data augmentation, and ensemble techniques addressing limited samples and noise. By combining the strengths of deep learning and weighted ensemble learning, WDRIV-Net potentially enhanced generalization. LID degeneration is a prevalent lumbar spinal condition impacting individuals globally, leading to reduced quality of life and significant socioeconomic burdens. Early

diagnosis is crucial, as delayed intervention risks disease progression, complicating treatment and increasing burdens on patients and healthcare systems. MRI imaging is the standard diagnostic tool, distinguishing comorbidity- and single-type degeneration to guide timely and appropriate treatment.

LID degeneration can progress rapidly under certain conditions, and distinguishing degeneration types is challenging due to factors like MRI image quality (e.g., anisotropic resolution), inter-class similarity, intra-class variation, and observation bias. Additionally, a shortage of physicians and the time-intensive nature of medical image analysis complicate diagnosis, underscoring the need for automated diagnostic systems. To address data limitations, we applied data augmentation to expand the sample size. Ensemble transfer learning, which combines automatic feature extraction through CNNs, deep learning, and knowledge transfer, is a widely used framework for medical image analysis, aiding in decision-making systems and LID application development. Integrating weighted ensemble learning, data augmentation, and transfer learning enables efficient and timely evaluation of lumbar spine abnormalities.

The major contribution of this research lies in developing WDRIV-Net, a non-equal-weight ensemble deep transfer learning framework that combines four pre-trained models (DenseNet169, ResNet101, InceptionV3, and VGG19) to classify LID degeneration types (single-type and comorbidity-type). WDRIV-Net achieved superior performance across all evaluation metrics, including accuracy, AUC, precision, recall, and F1-score, as shown in Table 2. It outperformed the baseline models, other widely studied models, and SOTA ensemble methods, delivering the best results in most cases. Specifically, WDRIV-Net achieved higher accuracy and AUC on the private test dataset compared to VGG19, InceptionV3, DenseNet169, ResNet101, and other SOTA ensemble methods. The improved performance, with lower false-positive and false-negative rates, demonstrates its potential to complement clinical decision-making effectively. This study shows that WDRIV-Net can handle classification tasks with size-constrained datasets and provides insights for clinical interventions and treatments for lumbar spine intervertebral disc degeneration.

This work has several limitations as well. First, the dataset used for training and testing included only sagittal-view images, excluding axial- and coronal-view data commonly used by medical experts for comprehensive diagnosis. The high-accuracy results of the experiments implied that the bias could be negligible or maintained at a small level for this dataset. Second, the original dataset was relatively small for the standard practice of deep learning, a challenge noted in other studies [16, 17, 67]. Future research could explore integrating multi-view data or designing models to merge information from different planes. Additionally, performance comparisons between transfer-learning-based models and non-transfer-learning approaches, such as ROI segmentation or bounding box cropping, could be evaluated when more data are available [67]. Third, the model determines whether an MRI exam reveals single or multiple symptoms of disc bulge, prolapse, or herniation, yielding an overall classification across spinal levels. However, it does not distinguish these symptoms at individual intervertebral disc levels, nor does it offer detailed symptom descriptions. Although the traits of disc levels of LID where available were reported in the medical text records and might be exploited for diagnosis, the information was not annotated directly in the image data. Future research

could address this limitation by designing more advanced models by fusing medical text records or using more detailed annotated image data.

Conclusion

In this study, we propose a non-equal-weighted transfer learning algorithm by ensembling four pre-trained baseline models for the automatic type classification of lumbar intervertebral disc bulge, prolapse, and herniation. The proposed model can be integrated into the medial assessment system for the initial screening of LID degeneration cases with differentiated symptoms where the priority of intervention/treatment is of concern. This potentially results in reduced medical costs, mitigates risks from diagnostic ambiguity, and accelerates treatment decisions for LID degeneration. The outcomes identified in this study will be beneficial for healthcare practitioners, physicians, patients, and researchers in medical image processing.

Methods

The experimental results are illustrated to test the efficiency and effectiveness of the proposed approach. A private MRI data set consisting of 1182 lumbar spine sagittal-view images was collected at the Fujian Medical University Union Hospital, China, and Pingtan Comprehensive Experimentation Area Hospital, Pingtan, China. We deployed TensorFlow and Keras as the Python toolkits in the conda environment to facilitate the training and testing procedures. A standard desktop computer using Windows 11 with 16 GB of RAM and an NVIDIA RTX A4000 graphical processor was used to evaluate the performance of models and the proposed ensemble approach.

MRI data set of lumbar spine degeneration

The private data set comprised 1182 MRIs of lumbar spine degeneration images in total collected at the Fujian Medical University Union Hospital, China, and Pingtan Comprehensive Experimentation Area Hospital, Pingtan, China, with 591 cases each of single- and comorbidity-type degeneration. We randomly divided the 591 cases into 551 cases of training data and 40 cases of testing data for both types (Table 1). The ratio of the training sample size to the verification sample size during training was set up to 80:20. The data were labeled and checked by three medical experts with more than 5 (expert #1), 10 (expert #2), and 20 (expert #3) years of clinical experience in medical image processing, respectively. The data were initially observed and labeled by expert #1 to identify the type of disc bulge, prolapse, or herniation by examining and recording the traits of the LID. Multi-view data (e.g., sagittal-, axial-, and coronal-view data) were merged comprehensively to facilitate the correct manual examination in cases where single- or partial-view data were not able to draw the final diagnosis. Further, in the inspection of sagittal-view data, the expert recorded the traits of disc levels of LID where available and used for subsequent diagnosis. In the output of manual inspection, the expert reported an additional flag showing the type information, and whether two or more of these symptoms presented. These first-stage results were cross-checked and corrected by expert #2 and thereafter by the expert #3. Potential inspection error outputs from expert #1 were corrected by expert #2, and the error outputs from expert #2 were updated by

expert #3 if identified in the sequential chain process. The examination process was conducted in compliance with the AAOS [68–70].

The final annotated dataset consisted of 1,182 cases with equal numbers of single- and comorbidity-type LID degeneration, forming the raw data for analysis. The deep learning model used representative sagittal-view data for training and testing. A retrospective review of patient cases from Fujian Medical University Union Hospital and Pingtan Comprehensive Experimentation Area Hospital was conducted for data collected between January and December 2022. MRI scans were obtained using Siemens 3 T MRI scanners, with cases involving multiple inspections excluded. The study was approved by the Ethics Committee of Fujian Medical University Union Hospital (2020YF023-01). Examples of the dataset are shown in Fig. 6.

Data pre-processing and data augmentation

The original MRI data resolution ranged from $320\times320\times3$ to $512\times512\times3$. To standardize, the images were resized to $224\times224\times3$ for input into DenseNet169, ResNet101, InceptionV3, and VGG19 [23, 24, 39, 40, 42, 55]. Images were categorized into single- or comorbidity-type groups. Pixel values were normalized to the range from 0 to 1, and resizing and shuffling generated standardized samples. The analysis was challenged by limited data. To partially solve this issue, we first exploited the publicly available larger dataset ImageNet to transfer learning the traits of the widely used data set and thereafter utilized a data-augmentation technique to enlarge the size of training data and ameliorate the disturbance from data noise [10, 11, 21]. This process generated approximately $\geq$12,000 augmented images, helping prevent overfitting. Figure 7 illustrates the data pre-processing and augmentation procedures.

Fine-tuning parameters and settings

An approach classifying the type of LID degeneration was proposed by using private MRI images of the lumbar spine. We utilized the data-augmentation technique by rotating (i.e., 20 degrees), width shifting (i.e., 0.05), height shifting (i.e., 0.05), and horizontally flipping the lumbar spine MRI images to solve the challenge of a limited sample.

(a) (b) (c)

Fig. 6 **a**–**c** Examples of lumbar spine MRI images of the single- and comorbidity-type degeneration in the collected private data set

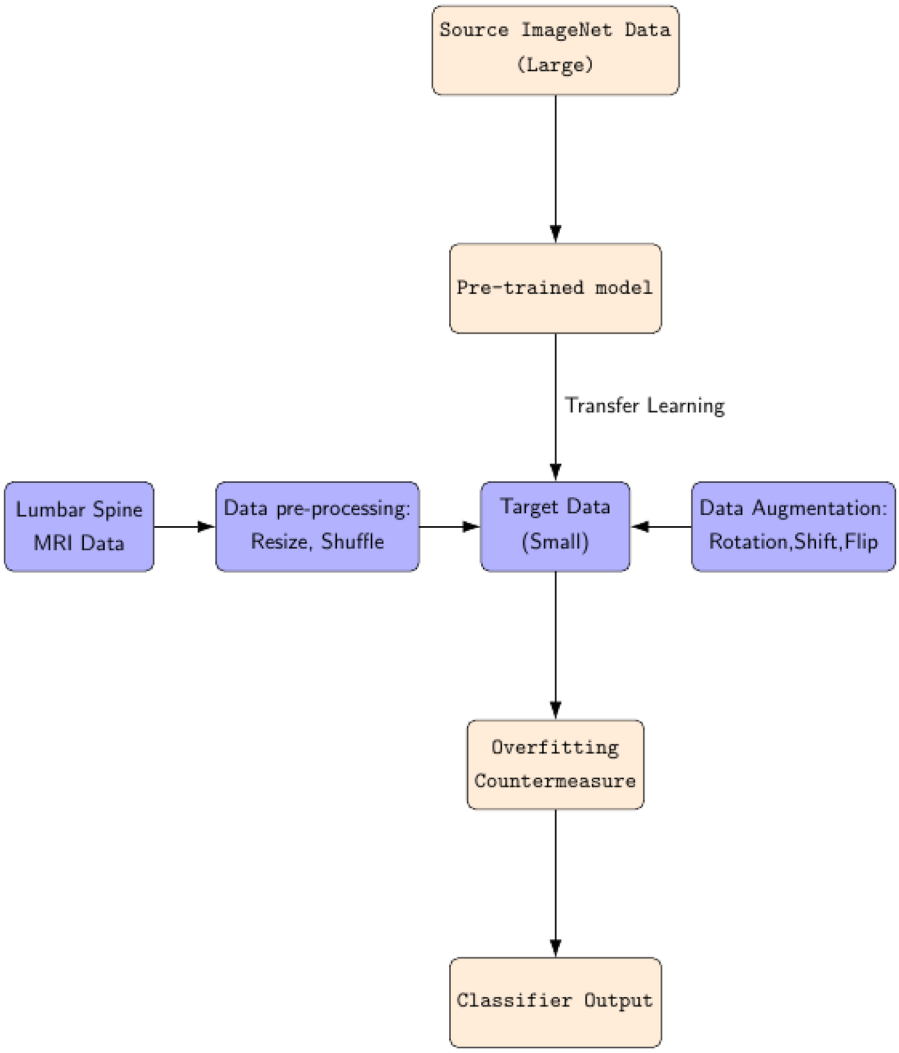

Fig. 7 Data pre-processing flow and data augmentation

Thereafter, a weighted ensemble-voting-based classification method was applied to predict the type of LID degeneration.

The proposed weighted ensemble deep transfer learning model was trained on the balanced and augmented dataset using categorical cross-entropy loss and the Adam optimizer (learning rate = 1e−4). Accuracy was estimated across both sub-categories, with training set for 25 epochs. An early stopping mechanism monitored performance, halting training after 10 epochs of no improvement to conserve computational resources. A dropout rate of 20% was used to mitigate overfitting. AUC values were also measured to evaluate the model's capacity to detect positive and negative cases.

Metrics for stratification

Accuracy, recall, precision, and F1-score

The benchmark models, the proposed approach, and other ensemble learning were examined using the test data set when the training procedure was completed by

exploiting the methodology of data augmentation and transfer learning. We evaluated the metrics including accuracy, recall, precision, F1-score, and AUC. We defined that true positive (TP) denotes the statistics information of comorbidity-type degeneration in the lumbar spine images, contrastingly true negative (TN) delineates the statistics of single-type degeneration. Consequently, false negative (FN) represents the statistics of comorbidity-type images incorrectly classified as single-type counterparts, and false positive (FP) demonstrates the statistics of single-type images that are incorrectly identified as comorbidity-type (Hashmi et al., 2020). The metrics of accuracy, precision, recall (or equivalently sensitivity), and F1 are determined by the formulas defined from Eq. (1) to (4):

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}, \tag{1}$$

$$Precision = \frac{TP}{TP + FP}, \tag{2}$$

$$Recall = \frac{TP}{TP + FN}, \tag{3}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{4}$$

ROC and AUC

Receiver operating characteristic (ROC) analysis is a standard methodology used to evaluate the performance of a binary classification system, which is applied extensively in clinical medicine analysis [43]. The ROC curve is a two-dimensional plotting that illustrates the relationship between the true and false-positive rates of a binary classifier.

AUC, which is identical to the probability that the decision value allocated to a randomly selected positive sample is greater than the value allocated to a negative sample (or equivalently, a randomly chosen negative case with a smaller estimated probability than a positive counterpart belonging to the positive type), is a univariate description of the ROC curve [43, 61]. The AUC curve is a performance metric sketching the extent to which a classifier divides types. Fundamentally, a classifier seeks the optimal formation of samples from a multi-dimensional feature space to a one-dimensional space during the training process. Hence, the range of AUC is typically larger than 0.5 and lower than 1, with greater values representing better performance as the area enclosed is greater. A proposed approach to calculate the AUC for the binary classification is illustrated in Eq. (5) [38]:

$$AUC = \frac{S_0 - \frac{n_0(n_0+1)}{2}}{n_0 n_1}, \tag{5}$$

where n_0 and n_1 denote, respectively, the statistics of positive and negative cases, and $S_0 = \sum r_i$, where r_i is the rank statistics of the ith positive case.

Metrics for weighted proposed ensemble model

The proposed ensemble model weighs each baseline model and the mathematical formulation is governed by the eq. from (6) to (7):

$$WPred = \prod \left(\sum_{i=1}^{4} \mathcal{W}_i Pred_i \right), \tag{6}$$

$$\sum_{i=1}^{4} \mathcal{W}_i = 1, \tag{7}$$

where $Pred_i (i = 1, 2, 3, 4)$ denotes the predicted outcome of the benchmark models, and $\mathcal{W}_i$ sketches the weights for the baseline models. The summation of all weights is equal to one. $\prod(\cdot)$ represents the algorithm of weighted ensemble deep learning, and *WPred* is the prediction outcome using the proposed weighted ensemble algorithm. In our analysis, $\mathcal{W}_i = \{0.2, 0.1, 0.3, 0.4\}$ for the benchmark resnet101, DenseNet169, InceptionV3, and VGG19, respectively. The weights for the models were calibrated and finally decided by partially referring to the performance and the loss observations of each model.

Benchmark models

An optimal approach to the automatic classifying of types of lumbar spinal complex degeneration was presented based on MRI images of LID. To solve the limited-data issue, we employed data augmentation and denoising techniques and exploited the ImageNet data set to pre-train four models and thereafter transferred the knowledge learned from ImageNet to the private MRI data set of LID based on weighted ensemble deep transfer learning, which enhanced the generalization of the models by integration of the merits of both deep learning, transfer learning, and weighted ensemble learning.

ImageNet data set

ImageNet is a data set accommodating more than 15 million high-resolution labeled images incorporating nearly 22,000 categories. The images in ImageNet are collected from the internet globally and labeled by professionals [62, 71]. Pre-trained models based on the ImageNet database are used to extract the features of medical images and have proven validity and effectiveness in predicting types, classification, segmentation, and other clinical outcomes [18, 32, 34–37, 62, 71].

Weighted transfer learning

Transfer learning is a widely used machine learning method in learning applicable knowledge from constrained data for prediction issues [16, 49, 51, 53, 64]. It has gained significant attention for its ability to transfer knowledge across different data distributions. The effectiveness of transfer learning has been validated in a variety of scenarios [16, 49, 51, 53, 63, 64], with four main models: instance-, feature-, parameter-, and relation-based [64].

A recent advancement, weighted ensemble deep transfer learning, combines deep learning, feature transfer learning, and weighted ensemble learning of individual

models [16, 63]. This approach enhances machine learning and transfer learning, offering better generalization, accuracy, flexibility, and robustness [16, 18, 48]. In our experiments, we observed that the non-equal-weight ensemble method outperformed the equal-weight approach.

Data augmentation

Recent breakthroughs in image processing have been driven by deep learning techniques, including classification [37, 39], object recognition [15, 21, 41], and segmentation [45, 46]. However, deep learning faces challenges such as small sample sizes and data quality. To address these, data augmentation generates synthetic data with alternative features while preserving the major traits of the original data, improving the performance of classifications [65, 66]. Additionally, combining data augmentation with transfer learning enhances model performance [59]. In this study, we employed both techniques to overcome these challenges.

VGG19 model

VGG stands for Visual Geometry Group, which is a deep-learning architecture with CNNs structures [55, 71]. VGG19, a variant of the VGG model, consists of 19 layers including 16 layers of CNNs, 3 fully connected layers (i.e., FC), and a softmax/sigmoid layer for prediction. VGG is a benchmark that outranks other deep neural baseline models in multiple vision tasks [55, 71, 72]. Studies exploring the effect of the CNNs depth on the accuracy in the sizable image-detection setting have identified that increasing depth to ≥ 19 weight layers with minuscule convolution filters can considerably enhance the performance of localization and classification, respectively. And the outcomes generalize well to other settings as well [15, 55].

The representative input to VGG typically consists of 224×224 RGB images with a receptive size of 3×3 by default [71]. A representative VGG19 model can comprise five blocks using CNNs and pooling layers to extract features, followed by FC layers and a softmax or sigmoid output layer for multi- or binary classification [55, 71]. Figure 8 outlines the schema of a benchmark VGG19 model used in the proposed ensemble analysis [55].

Resnet101 model

Resnet represents the residual network that reduces the cost of training [39, 40]. The "101" denotes the structure of weight layers. Resnet rephrases the networks as residual

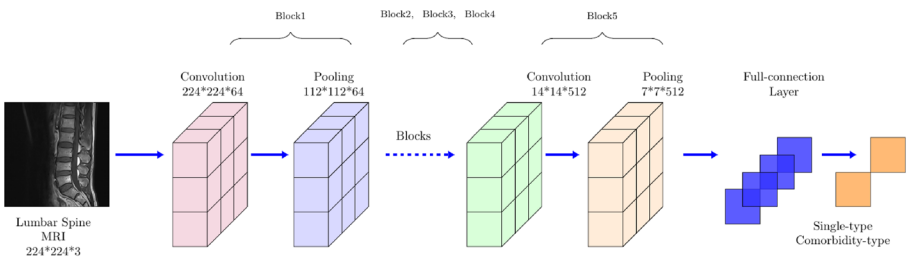

Fig. 8 Schema of a benchmark VGG19 model, and the compositional architecture used in the proposal

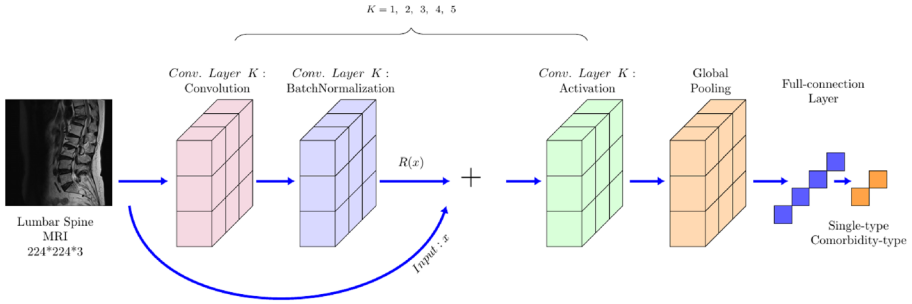

Fig. 9 Schema of a benchmark Resnet101 model, and the architecture used in the proposal analysis

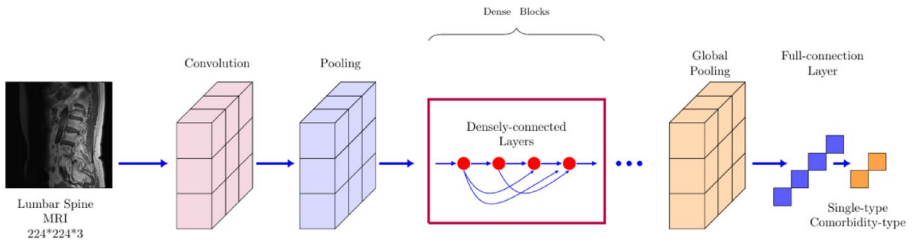

Fig. 10 Schema of a benchmark DenseNet169 model, and the compositional architecture used in the proposal appraisal

framework relative to the source input signal [39, 40, 73–75]. The Resnet framework exploits the identity shortcuts to capture essential residual information. Past studies have unveiled that the referenced residual networks excel in optimization with admissible accuracy [39, 40]. Experiments on ImageNet data set with a depth of over one hundred layers have yielded desirable outcomes with reduced complexity [39, 40, 74].

A typical ResNet model includes five layers of CNNs, each consisting of convolution, batch normalization, and activation processes. Figure 9 illustrates the ResNet101 model framework used in the proposed ensemble approach, where element-wise computation between the residual $R(x)$ and input x is conducted to better preserve features [39, 73, 74].

DenseNet169 model

DenseNet (Densely Connected Convolutional Networks) connects each layer to all preceding layers, using their feature maps as inputs for subsequent layers [42, 76, 77]. This architecture reduces issues like vanishing and exploding gradients, improves feature propagation, and reuses features while minimizing parameter size. DenseNet has shown significant performance improvements over SOTA networks with fewer computational resources. Typically, it features more layer connections compared to traditional CNNs. Figure 10 illustrates the DenseNet169 model architecture used in this study [42, 77].

InceptionV3 model

The Inception architecture optimizes local structures in a network using available components. InceptionV3 improves on InceptionV1 and V2 by addressing issues like label smoothing regularization and normalization, reducing overfitting. In our experiments, we used InceptionV3 with 1×1, 3×3, and 5×5 convolutions. We also tested

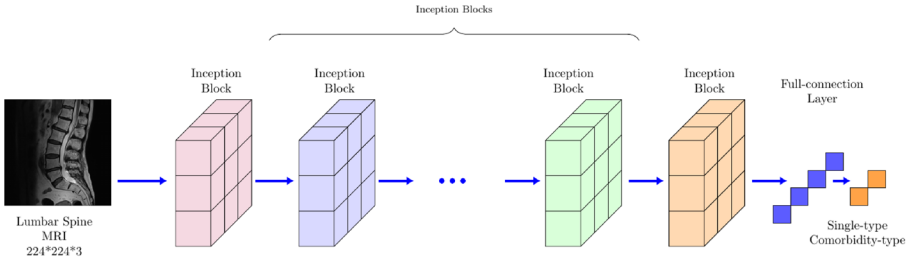

Fig. 11 Schema of a benchmark InceptionV3 model with 1 × 1, 3 × 3, and 5 × 5 convolutions, and the compositional architecture used in the analysis

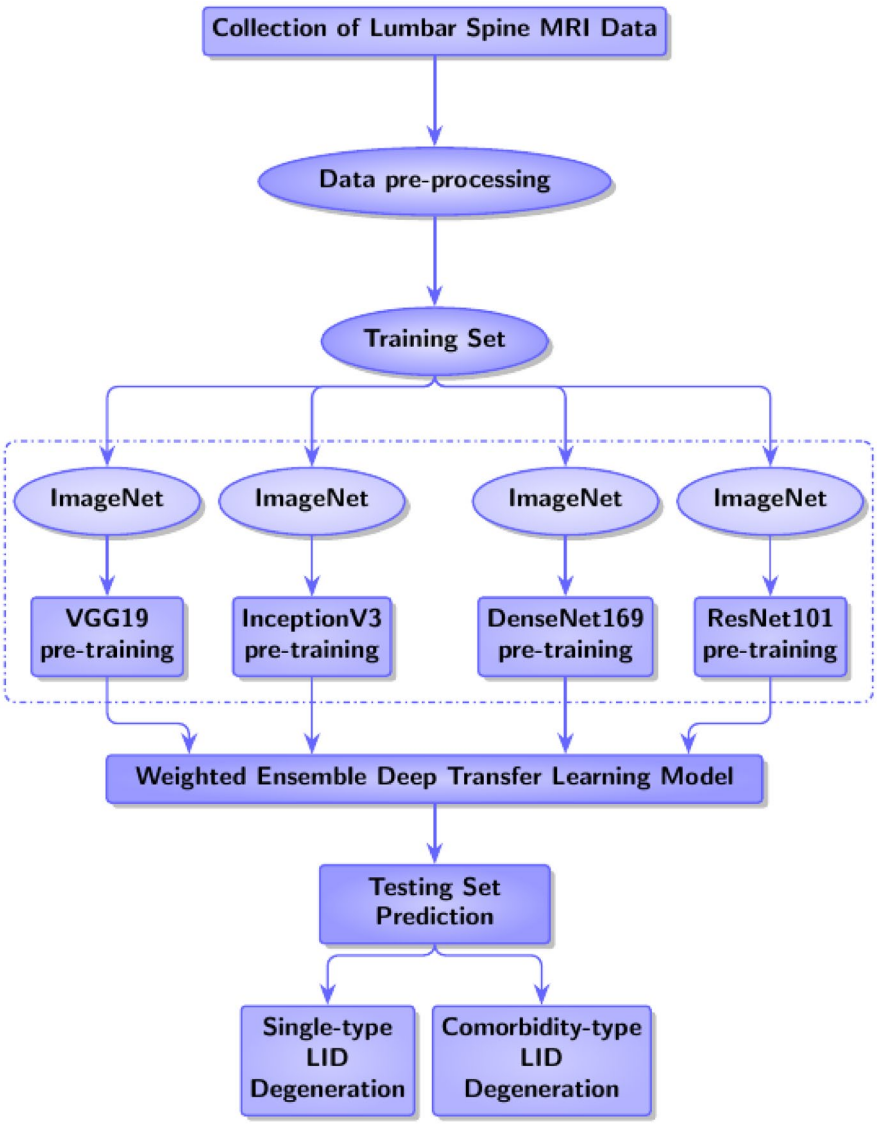

Fig. 12 The schema of the proposed ensemble deep transfer learning

InceptionResNetV2 with residual structures, but the ensemble framework with InceptionV3 outperformed InceptionResNetV2 in our setup (results not shown for brevity). Figure 11 illustrates the InceptionV3 model and its architecture used in our approach [23, 24].

The proposed weighted ensemble deep transfer learning framework

Ensemble deep transfer learning is a recent development that combines transfer learning and weighted ensemble learning techniques to generate a robust architecture yielding the consequential improvement of performance, as well as better generalization capacities over individual benchmark models [16].

The proposed WDRIV-Net framework uses non-equal-weight soft-ensemble voting to predict outcomes based on the probabilities from each baseline model. This approach combines the predictions of four pre-trained benchmark models (Densenet169, ResNet101, InceptionV3, VGG19) trained on the ImageNet dataset, improving overall performance and reliability while preventing gradient vanishing or exploding issues. The weighted soft-voting classifier stratifies predictions based on each model's probability, and an ensemble aggregator applies calibrated weights for final estimation. Figure 12 illustrates the WDRIV-Net framework and algorithm flow.

Abbreviations

AAOS	American Academy of Orthopaedic Surgeons
Adam	Adaptive moment estimation
AUC	Area under the curve
BBC	Bounding box cropping
CNNs	Convolutional Neural Networks
CT	Computerized tomography
DenseNet	Densely Connected Convolutional Networks
EL	Ensemble learning
EP	Endplates
FC	Full-connection layers
FN	False negative
FP	False positive
InceptionResNet	Inception Residual Networks
InceptionV1	Inception Networks Version 1
InceptionV2	Inception Networks Version 2
InceptionV3	Inception Networks Version 3
INP	Inner nucleus pulposus
LDB	Lumbar disc bulge
LDH	Lumbar disc herniation
LDP	Lumbar disc prolapse
LID	Lumbar intervertebral discs
MRI	Magnetic resonance imaging
OAF	Outer annulus fibrosus
ResNet	Residual Networks
ROC	Receiver operating characteristic
ROI	Region of interest
VGG	Visual Geometry Group
WDRIV-Net	The proposed non-equal-weight ensemble transfer learning framework integrating benchmarks including VGG19, InceptionV3, ResNet101, and DenseNet169
TL	Transfer learning
TN	True negative
TP	True positive

Acknowledgements

The authors thank Chunmei Chen for the helpful discussions and suggestions on the study.

Author contributions

Ichiro Nakamoto: conceptualization, investigation, methodology, programming, formal analysis, writing–original draft, revised version, review and editing, project administration, funding acquisition. Hua Chen: supervision, project administration, writing–review and editing, funding acquisition. Jianfeng Wu: investigation, data-processing, writing–review and

editing, funding acquisition. Rui Wang: resources, supervision, writing—review and editing. Yan Guo: Project administration, writing–review. Wei Chen: writing—review. Jie Feng: data-processing, investigation, writing—review.

Funding
This work was partially funded by the Pingtan Comprehensive Experimentation Area Hospital Technology project (grant no: Z2024001), Graduate School Publication Funding of Fujian University of Technology (grant no: YJC22-1), Fujian Province Zhilian Cloud Supply Chain Technology and Economic Integration Service Platform (grant no: KY310337), National Social Science Foundation of China (grant no: 22BGL007), and Humanities and Social Sciences Youth Funding of the Education Ministry of China (grant no: 24YJC790064).

Availability of data and materials
The data are confidential and the authors do not have permission to share the data.

Declarations

Ethics approval and consent to participate
The study was conducted in accordance with the Declaration of Helsinki. Approval for this study was granted by the Ethics Committee of Fujian Medical University Union Hospital, Fuzhou, China (2022KY007-1). Owing to the retrospective nature of the study and the minimal risks involved, patient informed consent was waived. The authors did not have access to information that could identify individual participants during or after data collection.

Consent for publication
The publication has been approved by all co-authors.

Competing interests
The authors declare no competing interests.

Received: 17 September 2023 Accepted: 20 January 2025
Published online: 06 February 2025

References

1. Wu Y, Wang Y, Wu J, Guan J, Mao N, Lu C, Lv R, Ding M, Shi Z, Cai B. Study of double-level degeneration of lower lumbar spines by finite element model. World Neurosurgery. 2016;86:294–9.
2. Rohlmann A, Zander T, Schmidt H, Wilke HJ, Bergmann G. Analysis of the influence of disc degeneration on the mechanical behaviour of a lumbar motion segment using the finite element method. J Biomech. 2006;39(13):2484–90.
3. Jiao M, Liu H, Yang Z, Tian S, Ouyang H, Li Y, Yuan Y, Liu J, Wang C, Lang N, Jiang L, Yuan H, Qian Y, Wang X. Self-supervised learning based on a pre-trained method for the subtype classification of spinal tumors. Comput Math Model Cancer Anal CMMCA. 2022;13574:58–67.
4. Kanna RM, Shetty AP, Rajasekaran S. Patterns of lumbar disc degeneration are different in degenerative disc disease and disc prolapse magnetic resonance imaging analysis of 224 patients. Spine J. 2014;14(2):300–7.
5. Raheem HM, Aljanabi M. Studying the bulging of a lumbar intervertebral disc: a finite element analysis. Procedia Struct Integr. 2020;28:1727–32.
6. Tang Y, Wu X, Ou-yang L, Li Z. An ambiguity-aware classifier of lumbar disc degeneration. Knowl-Based Syst. 2022;258: 109992.
7. Hussain M, Koundal D, Manhas J. Deep learning-based diagnosis of disc degenerative diseases using MRI: a comprehensive review. Comput Electr Eng. 2023;105: 108524.
8. Mousavi SM, Abdullah S, Niaki STA, Banihashemi S. An intelligent hybrid classification algorithm integrating fuzzy rule-based extraction and harmony search optimization: medical diagnosis applications. Knowl-Based Syst. 2021;220(23): 106943.
9. Morbée L, Chen M, Herregods N, Pullens P, Jans LNO. MRI-based synthetic CT of the lumbar spine: geometric measurements for surgery planning in comparison with CT. Eur J Radiol. 2021;144: 109999.
10. Miskin N, Gaviola GC, Huang RY, Kim CJ, Lee TC, Small KM, Wieschhoff GG, Mandell JC. Intra- and intersubspecialty variability in lumbar spine mri interpretation: a multireader study comparing musculoskeletal radiologists and neuroradiologists. Curr Probl Diagn Radiol. 2020;49(3):182–7.
11. Miskin N, Gaviola GC, Huang RY, Kim CJ, Lee TC, Small KM, Wieschhoff GG, Mandell JC. Standardized classification of lumbar spine degeneration on magnetic resonance imaging reduces intra- and inter-subspecialty variability. Curr Probl Diagn Radiol. 2022;51(4):491–6.
12. Jamaludin A, Kadir T, Zisserman A. SpineNet: automated classification and evidence visualization in spinal MRIs. Med Image Anal. 2017;41:63–73.
13. Cihangiroglu M, Yıldırım H, Bozgeyik Z, Senol U, Ozdemir H, Topsakal C, Yılmaz S. Observer variability based on the strength of MR scanners in the assessment of lumbar degenerative disc disease. Eur J Radiol. 2004;51(3):202–8.
14. Khare MR, Havaldar RH. Predicting the anterior slippage of vertebral lumbar spine using Densenet-201. Biomed Signal Process Control. 2023;86: 105115.
15. Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y. OverFeat: integrated Recognition, Localization and Detection using Convolutional Networks. In: Proc. 2014, ICLR.
16. Tavana P, Akraminia M, Koochari A, Bagherifard A. An efficient ensemble method for detecting spinal curvature type using deep transfer learning and soft voting classifier. Expert Syst Appl. 2023;213: 119290.

17. Tavana P, Akraminia M, Koochari A, Bagherifard A. Classification of spinal curvature types using radiography images: deep learning versus classical methods. Artif Intell Rev. 2023;56:13259–91.
18. Chen C, Zhang Q, Yu B, Yu Z, Lawrence PJ, Ma Q, Zhang Y. Improving protein–protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier. Comput Biol Med. 2020;123: 103899.
19. Chen Y, Lin Y, Xu X, Ding J, Li C, Zeng Y, Liu W, Xie W, Huang J. Classification of lungs infected COVID-19 images based on inception-ResNet. Comput Methods Programs Biomed. 2022;225: 107053.
20. Tanveer M, Rastogi A, Paliwal V, Ganaie MA, Malik AK, Ser JD, Lin CT. Ensemble deep learning in speech signal tasks: a review. Neurocomputing. 2023;550(14): 126436.
21. Soenksen LR, Kassis T, Conover ST, Marti-Fuster B, Birkenfeld JS, Tucker-Schwartz J, Naseem A, Stavert RR, Kim CC, Senna MM, Avils-Izquierdo J, Collins JJ, Barzilay R, Gray ML. Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. Science Translational Medicine. 2023;13(581):eabb3652.
22. Jean N, Burke M, Xie M, Davis WM, Lobell DB, Ermon S. Combining satellite imagery and machine learning to predict poverty. Science. 2016;253(6301):790–4.
23. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-ResNet and the impact of residual connections on Learning. arXiv. 2016. http://arxiv.org/abs/1602.07261.
24. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. arXiv. 2015. http://arxiv.org/abs/1512.00567.
25. Topal AO, Chitic R, Leprévost F. One evolutionary algorithm deceives humans and ten convolutional neural networks trained on ImageNet at image recognition. Appl Soft Comput. 2023;143: 110397.
26. Wang D, Lai A, Gansau J, Seifert AC, Munitz J, Zaheer K, Bhadouria N, Lee Y, Nasser P, Laudier D, Holguin N, Hecht AC, Latridis JC. Lumbar endplate microfracture injury induces Modic-like changes, intervertebral disc degeneration and spinal cord sensitization – an in vivo rat model. Spine J. 2023. https://doi.org/10.1016/j.spinee.2023.04.012.
27. Wang N, Yeung DY. Learning a deep compact image representation for visual tracking. Advances in neural information processing systems; 2013. p. 809–17.
28. Alanazi AH, Cradock A, Rainford L. Development of lumbar spine MRI referrals vetting models using machine learning and deep learning algorithms: comparison models vs healthcare professionals. Radiography. 2022;28:674–83.
29. Alves TS, Pinto MA, Venturac P, Nevesc CJ, Biron DG, Junior AC, Filho PLDP, Rodrigues PJ. Automatic detection and classification of honey bee comb cells using deep learning. Comput Electron Agric. 2020;170: 105244.
30. Avni U, Greenspan H, Konen E, Sharon M, Goldberger J. X-ray categorization and retrieval on the organ and pathology level, using patch-based visual words. IEEE Trans Med Imaging. 2010;30(3):733–46.
31. Balzer I, Mühlemann M, Jokeit M, Rawal IS, Snedeker JG, Farshad M, Widmer J. A deep learning pipeline for automatized assessment of spinal MRI. Comput Methods Progr Biomed Update. 2022;2: 100081.
32. Dar JA, Srivastava KK, Lone SA. Design and development of hybrid optimization enabled deep learning model for COVID-19 detection with comparative analysis with DCNN, BIAT-GRU. XGBoost Comput Biol Med. 2022;150: 106123.
33. Mar-Cupido R, García V, Rivera G, Sánchez JS. Deep transfer learning for the recognition of types of face masks as a core measure to prevent the transmission of COVID-19. Appl Soft Comput. 2022;125: 109207.
34. Douarre C, Schielein R, Frindel C, Gerth S, Rousseau D. Transfer learning from synthetic data applied to soil-root segmentation in X-ray tomography images. J Imaging. 2018;4(65):1–14.
35. Feng R, Zheng X, Gao T, Chen J, Wang W, Chen DZ, Wu J. Interactive few-shot learning: limited supervision, better medical image segmentation. IEEE Trans Med Imaging. 2021;40(10):2575–88.
36. Fujioka T, Yashima Y, Oyama J, Mori M, Kubota K, Katsuta L, Kimura K, Yamaga E, Oda G, Nakagawa T, Kitazumea Y, Tateishi U. Deep-learning approach with convolutional neural network for classification of maximum intensity projections of dynamic contrast-enhanced breast magnetic resonance imaging. Magn Reson Imaging. 2021;75:1–8.
37. Gayathri JL, Abraham B, Sujarani MS, Nair MS. A computer-aided diagnosis system for the classification of COVID-19 and non-COVID-19 pneumonia on chest X-ray images by integrating CNN with sparse autoencoder and feed forward neural network. Comput Biol Med. 2022;141: 105134.
38. Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach Learn. 2001;45:171–86.
39. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. arXiv; 2015. http://arxiv.org/abs/1502.01852.
40. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR); 2016. https://doi.org/10.1109/CVPR.2016.90.
41. Zhou Y, Liu Y, Chen Q, Gu G, Sui X. Automatic lumbar MRI detection and identification based on deep learning. J Digit Imaging. 2019;32:513–20.
42. Huang G, Liu Z, Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. arXiv; 2016. http://arxiv.org/abs/1608.06993.
43. Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng. 2005;17(3):299–310.
44. Huang J, Shen H, Wu J, Hu X, Zhu Z, Lv X, Liu Y, Wang Y. Spine Explorer: a deep learning based fully automated program for efficient and reliable quantifications of the vertebrae and discs on sagittal lumbar spine MR images. Spine J. 2020;20:590–9.
45. Huang M, Zhou S, Chen X, Lai H, Feng Q. Semi-supervised hybrid spine network for segmentation of spine MR images. Comput Med Imaging Graph. 2023;107: 102245.
46. Masood RF, Taj IA, Khan MB, Qureshi MA, Hassan T. Deep learning based vertebral body segmentation with extraction of spinal measurements and disorder disease classification. Biomed Signal Process Control. 2022;71: 103230.
47. Pandi SS, Senthilselvi A, Gitanjali J, ArivuSelvan K, Gopal J, Vellingiri J. Rice plant disease classification using dilated convolutional neural network with global average pooling. Ecol Model. 2022;474: 110166.
48. Niu W, Sun Y, Zhang X, Lu J, Liu H, Li Q, Mu Y. An ensemble transfer learning strategy for production prediction of shale gas wells. Energy. 2023;275: 127443.
49. Rahman M, Cao Y, Sun X, Li B, Hao Y. Deep pre-trained networks as a feature extractor with XGBoost to detect tuberculosis from chest X-ray. Comput Electr Eng. 2021;93: 107252.

50. Rahman T, Chowdhury MEH, Khandakar A, Islam KR, Islam KF, Mahbub ZB, Kadir MA, Kashem S. Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray. Appl Sci. 2020;10:3233.
51. Yang X, Zhang Y, Lv W, Wang D. Image recognition of wind turbine blade damage based on a deep learning model with transfer learning and an ensemble learning classifier. Renew Energy. 2021;163:386–97.
52. Zoph B, Vasudevan V, Shlens J, Le QV. Learning Transferable Architectures for Scalable Image Recognition. arXiv; 2017. http://arxiv.org/abs/1707.07012.
53. Zheng Y, Li C, Zhou X, Chen H, Xu H, Li Y, Zhang H, Li X, Sun H, Huang X, Grzegorzek M. Application of transfer learning and ensemble learning in image-level classification for breast histopathology. Intell Med. 2023;3:115–28.
54. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: inverted residuals and linear bottlenecks. arXiv; 2018. http://arxiv.org/abs/1801.04381.
55. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv; 2014. http://arxiv.org/abs/1409.1556.
56. Hashmi S, Vanderwert RE, Price HA, Gerson SA. Exploring the benefits of doll play through neuroscience. Front Hum Neurosci. 2020;14: 560176.
57. Hoffmann J, Bar-Sinai Y, Lee LM, Andrejevic J, Mishra S, Rubinstein SM, Rycroft CH. Machine learning in a data-limited regime: augmenting experiments with synthetic data uncovers order in crumpled sheets. Sci Adv. 2019;5:eaau6792.
58. Isensee F, Schell M, Pflueger I, Brugnara G, Bonekamp D, Neuberger U, Wick A, Schlemmer HP, Heiland S, Wick W, Bendszus M, Maier-Hein KH, Kickingereder P. Automated brain extraction of multisequence MRI using artificial neural networks. Hum Brain Mapp. 2019;40(17):4952–64.
59. Jahja HD, Yudistira N. Mask usage recognition using vision transformer with transfer learning and data augmentation. Intell Syst Appl. 2023;17: 200186.
60. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science. 2015;349(6245):255–60.
61. Lee WH, Gader PD, Wilson JN. Optimizing the area under a receiver operating characteristic curve with application to landmine detection. IEEE Trans Geosci Remote Sens. 2007;45(2):389–97.
62. Li X, Cen M, Xu J, Zhang H, Xu XS. Improving feature extraction from histopathological images through a fine-tuning ImageNet model. J Pathol Inf. 2022;13: 100115.
63. Luo R, Bocklitz T. A systematic study of transfer learning for colorectal cancer detection. Inf Med Unlocked. 2023;40: 101292.
64. Xu H, Li W, Cai Z. Analysis on methods to effectively improve transfer learning performance. Theoret Comput Sci. 2023;940:90–107.
65. Zhang AS, Xu A, Ansari K, Hardacker K, Anderson G, Alsoof D, Daniels AH. Lumbar disc herniation: diagnosis and management. Am J Med. 2023;136(7):645–51.
66. Zhang M, Wang Z, Wang X, Gong M, Wu Y, Li H. Features kept generative adversarial network data augmentation strategy for hyperspectral image classification. Pattern Recogn. 2023;142: 109701.
67. Šušteršič T, Ranković VV, Milovanović V, Kovačević V, Rasulić L, Filipović N. A deep learning model for automatic detection and classification of disc herniation in magnetic resonance images. IEEE J Biomed Health Inform. 2022;26(12):6036–46. https://doi.org/10.1109/JBHI.2022.3209585.
68. American Academy of Orthopaedic Surgeons (AAOS). https://www.aaos.org/.
69. Fardon DF, Milette PC. Nomenclature and classification of lumbar disc pathology. Recommendations of the Combined Task Forces of the North American Spine Society, American Society of Spine Radiology, and American Society of Neuroradiology. Spine (Phila Pa 1976). 2001;26(5):E93–113.
70. Truumees E, Prather H, editors. Orthopaedic knowledge update: spine 5. American Academy of Orthopaedic Surgeons; 2017, Chapter 17: Lumbar Disk Herniations.
71. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. NIPS; 2012. p. 1106–14.
72. Dean J, Corrado G, Monga R, Chen K, Devin M, Mao M, Ranzato M, Senior A, Tucker P, Yang K, Le QV, Ng AY. Large scale distributed deep networks. NIPS; 2012. p. 1232–40.
73. Bishop CM. Neural networks for pattern recognition. Oxford University Press; 1995.
74. Ripley BD. Pattern recognition and neural networks. Cambridge University Press; 1996.
75. Venables W, Ripley B. Modern applied statistics with s-plus. Springer Science & Business Media; 1999.
76. Fahlman SE, Lebiere C. The cascade-correlation learning architecture. NIPS; 1989. p. 2.
77. Wilamowski BM, Yu H. Neural network learning without backpropagation. IEEE Trans Neural Netw. 2010;21(11):1793–803.

Publisher's Note