Journal of Scientific & Industrial Research
Vol. 82, August 2023, pp. 818-830
DOI: 10.56042/jsir.v82i08.3396

Rank Based Two Stage Semi-Supervised Deep Learning Model for X-Ray Images Classification

Pawan Kumar Mall[1]*, Vipul Narayan[2]*, Swapnita Srivastava[2], Munish Sabarwal[2], Vimal Kumar[2], Shashank Awasthi[3,4] & Lalit Tyagi[3]

[1]Madan Mohan Malaviya University of Technology, Gorakhpur 273 016, Uttar Pradesh, India
[2]Galgotias University, Greater Noida 203 201, Uttar Pradesh, India
[3]GL Bajaj Institute of Technology and Management, Greater Noida203 201, Uttar Pradesh, India
[4]Research Management Center, Management and Science University, 40100 Shah Alam, Selangor Darul Ehsan, Malaysia

Received 23 August 2021; revised 07 March 2023; accepted 29 June 2023

Deep learning approaches rely on a wide-scale labeled dataset to attain a high level of performance. Although labeled data is more difficult and costly to access in some applications, such as bioinformatics and medical imaging, wide variety of ongoing research on the topic of Semi-Supervised Deep Learning (SSDL) can improve and fix underlying problems in this domain. The motivation for the suggested model Rank Based Two-Stage Semi-Supervised Deep Learning (RTS-SS-DL) is the same as how doctors deal with unobserved or suspect cases in day to day practice. The physicians deal with these suspect instances with the help of professional assistance from their colleagues. Before beginning therapy, some patients seek the opinion of a variety of skilled professionals. The patients are treated by the most appropriate (vote count) professional diagnosis. Our model (RTS-SS-DL) has achieved impressive metrics including 92.776% accuracy, 97.376% specificity, 86.932% sensitivity, 96.192% precision, 85.644% MCC (Matthews Correlation Coefficient), 3.808% FDR (False Discovery Rate), 2.624% FPR (False Positive Rate), 91.072% f1-score, 90.85% NPV (Negative Predictive Value), and 13.068% FNR (False Negative Rate) for the unseen dataset. The outcome of this research results in an SSDL model that is both more precise and effective.

Keywords: Labeled dataset, RTS-SS-DL, Self-organising classifier, Semi-supervised learning, Shoulder's fracture classification

Introduction

In today's era, Semi-Supervised-Deep-Learning (SS-DL) has been identified as a promising new research pathway in the computer vision domain. The notion of SS-DL initially emerged in the 1970s.[1–3] the motivation behind using SS-DL in medical dataset is to enhance the efficiency and accuracy of machine learning models used for medical image analysis. Medical imaging is a critical area of healthcare, with imaging techniques such as X-rays, MRI scans, and CT scans used to diagnose and monitor a wide range of diseases and conditions. However, obtaining large amounts of labeled medical image data can be challenging, as it often requires expert annotation and can be time-consuming and expensive.[4] This can limit the effectiveness of supervised learning approaches that rely on large labeled datasets. SS-DL provides a way to leverage both labeled and unlabeled medical image data to train machine learning models. By using unlabeled data in conjunction with labeled data, the model can learn more generalize-able features that are useful for detecting patterns and abnormalities in medical images. This can lead to better accuracy and generalize-ability of the models, as well as reducing the amount of labeled data needed for training. Overall, SS-DL has the potential to enhance the efficiency and effectiveness of medical imaging analysis, ultimately leading to better patient outcomes and more efficient use of healthcare resources. With regard to medical imaging tasks, Semi-Supervised Learning (SSL) may use unlabeled data to enhance model performance. However, pseudo-labeling-based semi-supervised approaches have two issues with medical image datasets: the models' predictions are biased towards the majority class in unbalanced datasets, and the loss of useful information occurs

*Author for Correspondence
E-mail: pawankumar.mall@gmail.com,
vipulupsainian2470@gmail.com

when unlabeled data with confidence below the thresholds are discarded. FullMatch is a unique SSL architecture that utilizes all unlabeled input to improve the model's performance in order to address these problems.[6] The prominent SSDL research works and its application implementation in Deep Learning Model (DLM) minimize the shortage of labeled data need in the quest for a more data-efficient deep learning approach. Semi-supervised learning can be a powerful approach in medical imaging, as it can help to overcome the challenges of limited labeled data and enhance the accuracy and generalize-ability of models. However, it is important to carefully evaluate the performance of the model on both the labeled and unlabeled data to ensure that it is generalizing well and not simply over-fitting to the labeled data. The SSDL framework is shown in Fig. 1.

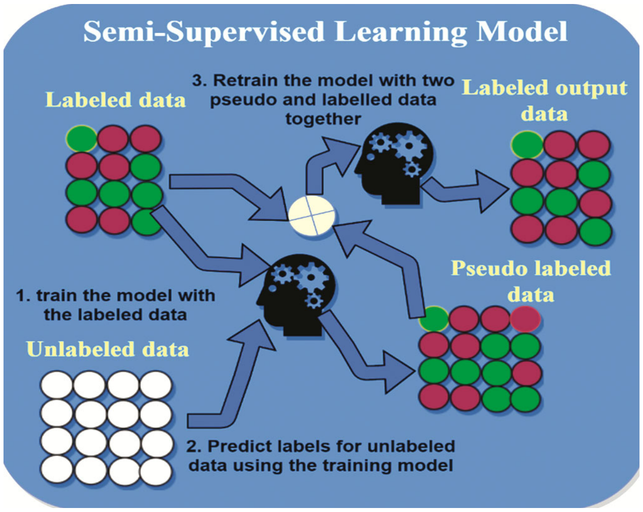

Semi-Supervised Learning Model
Labeled data
3. Retrain the model with two pseudo and labelled data together
Labeled output data
Pseudo labeled data
1. train the model with the labeled data
Unlabeled data
2. Predict labels for unlabeled data using the training model

Fig. 1 — Details of semi supervised deep learning

We have introduce a new rank-based two-stage SSDL model. The suggested (RTS-SSDL) framework enhance SS-DL models performance. The research findings demonstrate that the RTS-SSDL model performs significantly better than the conventional models and produces sufficient classification results. A comparison chart of different ensemble learning techniques is given in Table 1.

It's important to note that the strengths and weaknesses listed above are generalizations and may vary depending on the specific implementation and dataset used. The choice of ensemble learning technique should always be based on careful experimentation and evaluation.

We have proposes a novel SS-DL model for the classification of X-ray images. The major finding of the article can be given as follows:

Rank-based semi-supervised learning: The paper introduces a novel rank-based semi-supervised learning approach that the standard DLM are re-trained with the both pseudo dataset by each models. The model ranks the labeled samples and utilizes them to supervise the model's training, while the unlabeled samples are used to improve the model's representation learning.

Two-stage learning: The suggested model consists of two stages: pre-training and fine-tuning. The pre-training stage uses the unsupervised learning method to learn the underlying representation of the X-ray images. The fine-tuning stage utilizes both labeled and unlabeled samples to fine-tune the pre-trained model for classification.

The future scope of the presented model includes its application to other medical imaging tasks, such as

Table 1 — Comparison chart of different ensemble learning techniques

Ensemble Learning Technique	Base Models	Strengths	Weaknesses
Bagging	Any model	Reduces variance and overfitting, improves accuracy, works well with unstable models like decision trees	Can be computationally expensive
Boosting	Weak models	Can improve accuracy of weak models, works well with high bias and low variance models like decision stumps	Can overfit and be computationally expensive
Stacking	Any model	Can improve accuracy of individual models, works well with models that have complementary strengths and weaknesses	Can be computationally expensive
Random Forest	Decision trees	Reduces variance and overfitting, works well with datasets with many features	Can be computationally expensive
Gradient Boosting	Weak models	Can improve accuracy of weak models, works well with datasets with few features	Can overfit and be computationally expensive
AdaBoost	Weak models	Can improve accuracy of weak models, adapts to misclassified samples	Can overfit and be computationally expensive
Voting	Any model	Simple and easy to implement, can work well with models that have different strengths and weaknesses	May not improve accuracy if all models are similar

CT and MRI image classification. Additionally, the suggested rank-based SS-DL approach can be extended to other DLM and applications beyond medical imaging, where labeled data is scarce or expensive to obtain. The model's architecture and training methodology can also be further optimized to improve its scalability and efficiency for real-world deployment.

Related Work

SSDL is a well-known research area in medical images that aims to develop techniques to leverage unlabeled data for training DLM. SSDL has gained significant attention in recent years due to the growing need for large labeled datasets, which are often time-consuming and expensive to obtain.[5–7] There are several popular SSDL techniques that have been developed, including:

Self-Training: In this technique, a model is trained on labeled data and then used to generate predicted labels for unlabeled data.[6–9] The predicted labels are treated as if they were true labels, and the model is retrained on the labeled and pseudo-labeled data.[8–11]

Co-Training: This technique is used when there are two or more modalities of data available. Two separate models are trained on different modalities of the data, and each model is used to generate predicted labels for the other modality.[10] The predicted labels are then used as if they were true labels for the other modality, and the models are retrained on the labeled and pseudo-labeled data.[12]

Semi-Supervised Generative Adversarial Networks (GANs): GANs are a type of deep learning model that can be trained on both labeled and unlabeled data. The generator part of the GAN is trained on the unlabeled data to generate synthetic data that is similar to the labeled data. The discriminator part of the GAN is trained on both the labeled and synthetic data to distinguish between the two.[13]

The use of SSDL techniques in DLM has been shown to be effective in alleviating the requirement for labeled data. The SSDL has been successfully applied in a wide range of applications, including image recognition, speech recognition, and natural language processing.[14]

Theoretical Considerations for the Suggested Model

Dataset

One of the biggest collections of medical X-rays data of the bones is the musculoskeletal radiograph dataset. The dataset includes X-rays from January 2014 to December 2017 covering a period of four years at hospital, as well as a total of 58817 pictures from 21456 radiography case studies. The patients had an average age of 7.2 years, and 57% of them were male. Total X-ray images of the bones total 40561 in the (MURA) musculoskeletal radiograph. The collection contains 55.63% normal and 44.36% abnormal X-ray images. From the MURA dataset, we only took into account the shoulder study; as a result, the new dataset is now known as MURA-SU for our experiment 8942 train set and 194 test set. HATA-SU dataset is collected from HATA CHC train set contains 598 records and unseen 681 records for the test set, under supervision of Dr. Prashant Kumar Mall, Dr. Richa Singh and Dr. Siddharth Jaiswal from Kushinagar district. Both the dataset are detailed in Table 2.

Table 2 — The MURA-SU X-ray and HATA-SU dataset details

Dataset	Train Set	Test Set
MURA-SU	8942 (Normal Abnormal)	194 (Normal Abnormal)
HATA-SU (unlabeled)	598	0
HATA-SU (unseen)	0	681(Normal-381, Abnormal-300)
Total Size:10415		

Deep Learning Benchmark Models

The following is a succinct description of the key technical elements of the benchmark DLM as depicted in Fig. 2:

a) MobileNet

Mobilenet is a lightweight model.[15] The model was developed to resolves few hardware-related challenges, including power, energy, and limited memory. The model works on the concept of depth-wise separable convolutions.

b) Pre-Act ResNet

The pre-activation resnet model[16] Pre-Act ResNet is a type of DNN architecture used for image classification tasks. It is an improvement over the traditional ResNet architecture, which was introduced in 2015 by researchers at Microsoft. In Pre-Act ResNet, the order of batch normalization and activation functions is changed compared to the traditional ResNet architecture. Specifically, in Pre-Act ResNet, batch normalization is applied before the activation function, whereas in traditional ResNet, the activation function is applied before batch

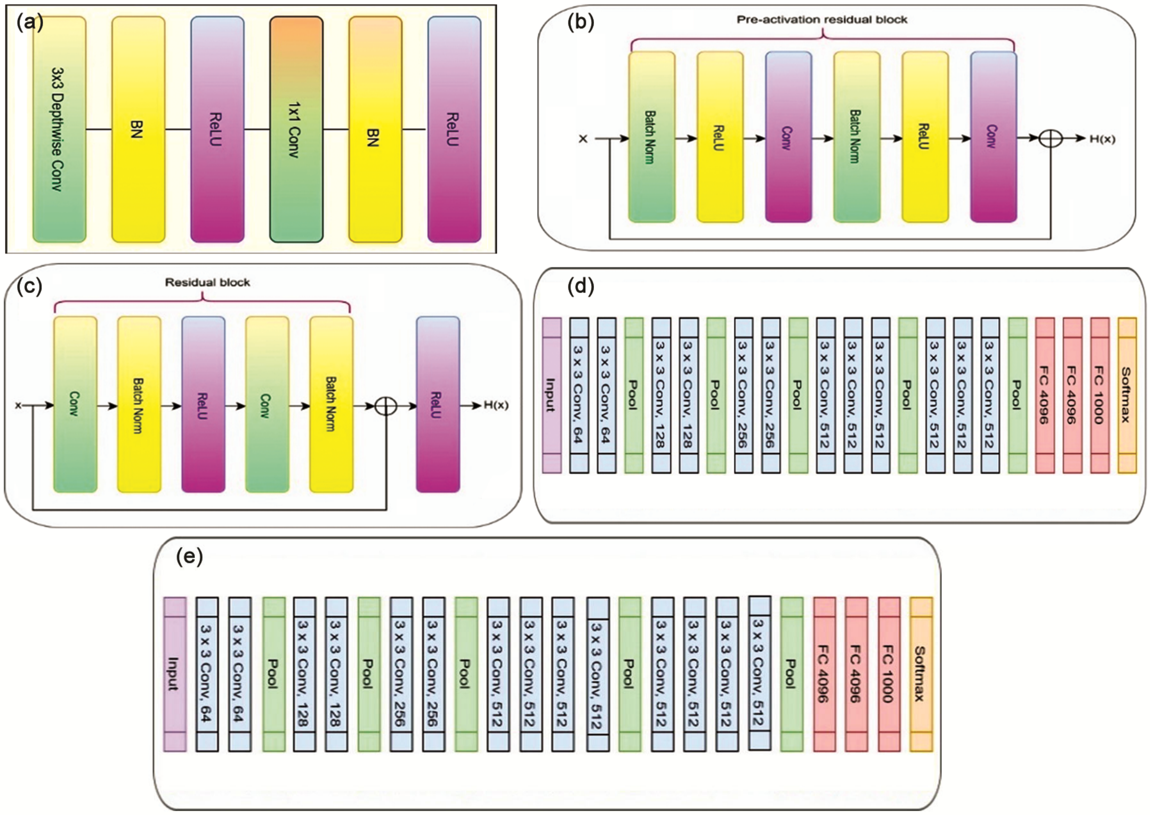


Fig. 2 — Structure of: (a) MobileNet, (b) Pre-Act ResNet, (c) ResNet-18, (d) VGG-16, (e) VGG-19

normalization. The main benefit of this modification is that it allows for better gradient flow during training. By applying batch normalization before the activation function, the normalization process is less likely to "wash out" the signal from the activation function, resulting in better learning and improved accuracy. Pre-Act ResNet has been shown to outperform traditional ResNet architectures on a number of benchmark datasets for image classification, including CIFAR-10, CIFAR-100, and ImageNet.

c) ResNet18

The ResNet-18[(17)] is a 18 layers deep. The ResNet design was devised, along with the concept of "skip connections." Residual connections allow parameter gradients to propagate relatively smoothly from the output layer to the network's prior layers, facilitates a train of more deep networks.

d) VGG-16

VGG-16[(18)] VGG-16 is an architecture that comprises 16 layers. VGG-16 is renowned for its uncomplicated and graceful design, as well as its exceptional performance in recognizing images. It set the highest benchmark in a competition that gauges the performance of computer vision models on a large-scale image dataset.

e) VGG-19

VGG-19[(19)] has been widely used as a benchmark model in computer vision tasks, such as image classification, object detection, and image segmentation. The network's pertained weights have been made publicly available and have served as a starting point for many other deep learning applications.[20]

Proposed Work

In the medical field, diagnosing patients using costly equipment and the input of multiple healthcare professionals to label data is a costly process. Our suggested model has six different stages: pre-processing images, determining rank, creating the model, generating a pseudo dataset, retraining DLM, and evaluating performance. The most important aspect of our framework is determining the rank of the benchmark DLM and retraining it with both labeled and pseudo datasets. Finally, the effectiveness of the suggested model is validated using the HATA-SU dataset, which has not been previously used. The working of the suggested framework is illustrated in Fig. 3.

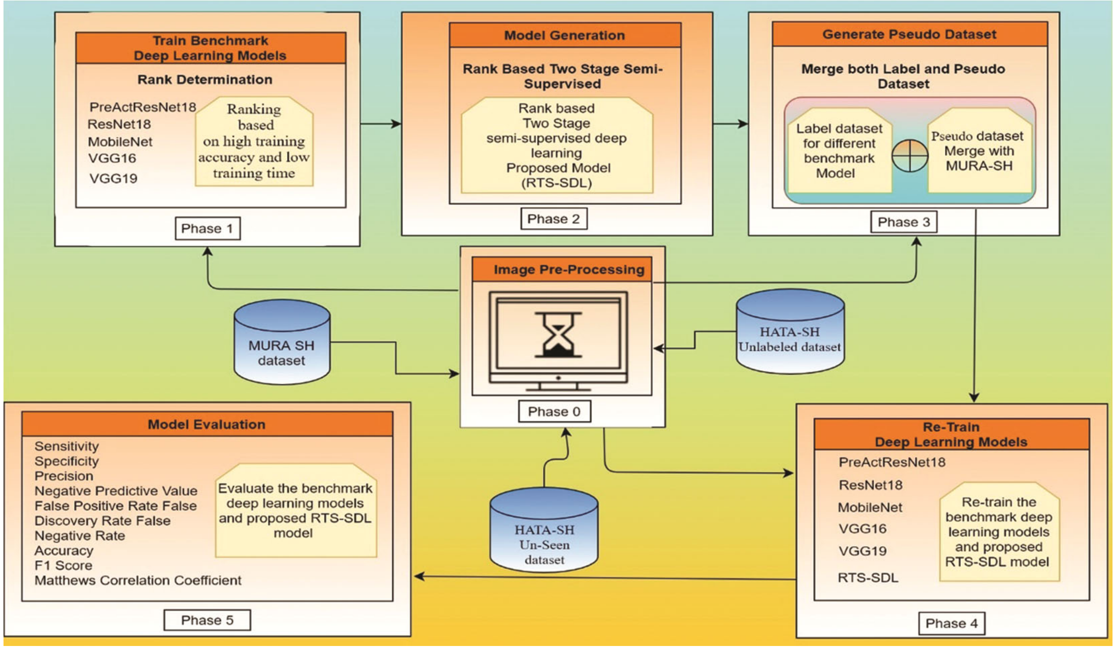

Train Benchmark
Deep Learning Models
Rank Determination
PreActResNet18
ResNet18
MobileNet
VGG16
VGG19
Ranking based on high training accuracy and low training time
Phase 1
Model Generation
Rank Based Two Stage Semi-Supervised
Rank based Two Stage semi-supervised deep learning Proposed Model (RTS-SDL)
Phase 2
Generate Pseudo Dataset
Merge both Label and Pseudo Dataset
Label dataset for different benchmark Model
Pseudo dataset Merge with MURA-SH
Phase 3
Image Pre-Processing
Phase 0
MURA SH dataset
HATA-SH Unlabeled dataset
Model Evaluation
Sensitivity
Specificity
Precision
Negative Predictive Value
False Positive Rate False
Discovery Rate False
Negative Rate
Accuracy
F1 Score
Matthews Correlation Coefficient
Evaluate the benchmark deep learning models and proposed RTS-SDL model
Phase 5
HATA-SH Un-Seen dataset
Re-Train
Deep Learning Models
PreActResNet18
ResNet18
MobileNet
VGG16
VGG19
RTS-SDL
Re-train the benchmark deep learning models and proposed RTS-SDL model
Phase 4

Fig. 3 — Diagram illustrating the working of suggested framework

Research Environment

The investigation was conducted in a digital setting utilizing a virtual machine. The primary virtual machine runs on an Ubuntu operating system, with a RAM of 14 GB and ten virtual CPUs derived from the AMD Ryzen 7 4800H GHz processor. The suggested model was implemented using Python 3.0.

Image Preprocessing

Pre-processing X-ray image normalization is the process of scaling pixel values to a standardized range. It helps to eliminate variations in illumination conditions and ensures that images have consistent intensity distributions. Common normalization methods include min-max scaling and z-score normalization. Images are often resized or scaled to a standard size or resolution to ensure consistency in further analysis. This can be achieved by resampling the image into 64×64.

Rank Determination

This stage in determining DLM rankings is crucial. Our suggested ranking method concentrates on elements of various DLM' training accuracy (Train_acc), test accuracy (Test_acc), and training time (Elsp_train_time). A threshold-based ranking algorithm is used to establish the model's rating.

Algorithm 1: Procedure for the Threshold based Ranking among the Standard DLM:

```
INPUT: $\mathrm{Train_{acc}}$, $\mathrm{Test_{acc}}$, $\mathrm{Elsp_{train_{time}}}$, $\mathrm{RankTrain_{acc}}$, $\mathrm{RankTest_{acc}}$, $\mathrm{RankElsp_{train\_}time}$, $N$

OUTPUT: Rank[ ]

Initialization;
Train_acc = Train accuracy,
Test_acc= Test accuracy,
Elsp_train_time = Training time elapsed,
RankTrain_acc = Rank Train accuracy,
RankTest_acc = Rank Test accuracy
$Rank_{Elsp\_Traintime}$ = Rank Training time elapsed
$N$ = Number of deep learning model
Rank[ ]= Rank assign to deep learning model
$\alpha = \sum_{i=1}^{N} Train_{Acc}/N$ * Compute train threshold*/
$\beta = \sum_{i=1}^{N} Test_{Acc}/N$ /* Compute test threshold*/
$\gamma = \sum_{i=1}^{N} Elsp_{Traintime}/N$ /* Compute elapsed threshold*/
For i=1 to N do /* Compute rank for N deep learning model*/
    If $Elsp_{Traintime} < \gamma$/* Compare Elasp train time with threshold*/
        Rank[i] = $Rank_{Elsp\_Traintime}[i]$/* Assign elapsed train time rank to model rank*/
    If $Test_{Acc} > \beta$ /* Compare test accuracy with test threshold */
        Rank[i] = $Rank_{test\_Acc}[i]$ /* Assign test rank to model rank*/
    If $Train_{Acc} > \alpha$ /* Compare train accuracy with train threshold*/
```

```
10. Rank[i] = $Rank_{Train\_Acc}[i]$ /* Assign train rank to model rank*/
11. Else
12. Rank[i] = $Rank_{Elsp\_Traintime}[i]$ /* Assign elapsed train time rank to model rank*/
13. End For
    Note: In case of tie between ranking priority is $Rank_{Elsp\_Traintime}$, $RankTrain_{Acc}$, $RankTest_{Acc}$
```

Model Generation

The model generation stage is crucial part of our suggested model. The inspiration behind the suggested model (RTS-SSDL) is the same as how doctors deal with the unobserved cases in the day to day life. The doctors tackle these unseen cases based on expert advice from their colleagues. Some patients seek the advice of several professional specialists before starting the treatment. The patients follow treatment according to maximum (vote count) expert diagnoses. In this study, five standards DLM are considered for the experiment. The suggested algorithm 2 is used for the two stage rank based model generation model. The detailed proposed structure is shown in Fig. 4.

Algorithm 2: Procedure for the two Stage Rank based Model Generation

```
INPUT: $Ud$, N, M_Rank [], M
OUTPUT:$Pdpm$
Initialization;
$Ud$ = Unlabled dataset,
$N$ = Number of Standard Deep learning model,
M_Rank[] =Array of Standard Deep learning modelssorted according to ascending order,
$M$ = Size of unlable dataset,
fusion_classifier_Rank =Deep learning models
MDM= Master DL model
$Pd$_MDM = Pseudo_dataset
GENERATE_PSEUDO_LABEL () method to generate pseudo label
1. For k=1 to 3 do /* Compute Pseudo dataset for three models*/
   2. Fusion_classifier_Rank[k] ={M_Rank[k], M_Rank [4], …… M_Rank[N]} /*Generate Three sub model according to top three Rank */
   3. $Pd$_Fusion_classifier ← GENERATE_PSEUDO_LABEL ($Ud$, Fusion _ classifier_Rank[k])
4. End for
5. MDM [] =logisticRegression {Fusion_classifier_Rank [1], Fusion_classifier_Rank [2], Fusion_classifier_Rank [3]}
6. $Pd$_ MFC ← GENERATE_PSEUDO _LABEL ($Ud$, MFC []}
```

Generate Pseudo Dataset

This step aims to label a pseudo medical data for the unlabeled HATA-SU data. The pseudo data are label based on the vote total, and the label with the highest votes serves as the pseudo label for both suggested and traditional models. The fake dataset is produced using technique 3 in a step-by-step fashion.

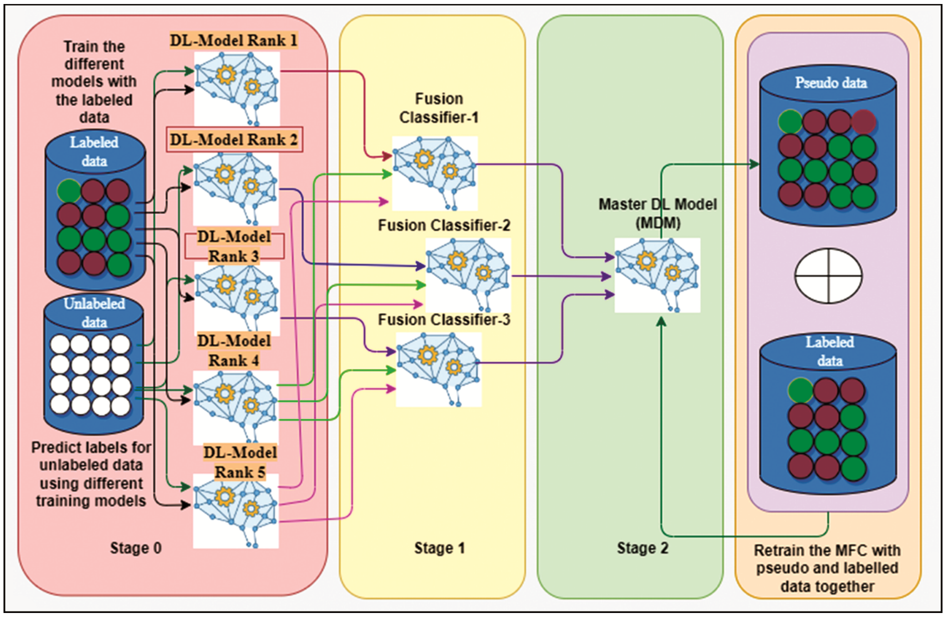


Fig. 4— Suggested semi-supervised rank based two stage deep learning model

Algorithm 3: Procedure for Pseudo Label Generation for Unlabeled Dataset

```
INPUT: $Ud$, N, M_Rank [], M
OUTPUT:$Pdpm$
Initialization;
$Ud$ = Unlabled dataset,
$N$ = Number of Standard Deep learning model,
M_Rank[]=Array of Standard Deep learning models sorted according to ascending order,
$M$ = Size of unlable dataset,
Fusion_classifier_Rank =Set of deep learning models
$Pdpm$ = Pseudo dataset,
GENERATE_PSEUDO_LABEL ($Ud$, Fusion_classifier_Rank [])
1.  M          $Class2$
    End for
```

Re-train the DML with Pseudo Data and MURA-SU

At this stage, the conventional DLM is re-trained using both the pseudo dataset and the MURA-SU (RTS-SS-DL) proposed model.

Validation of suggested Model

The effectiveness of the (RTS-SSDL) suggested model is evaluated and validated in the final stage of the experiment using an unseen dataset called HATA-SU.

Evaluation and Validation

The suggested framework assessment and validation are detailed in the following sections.

DML Evaluations

This tool is a popular Statistical measure, which generates tabular reports of the number of wrong and right predictions. The four key terminologies related to confusion matrix are TP (True Positive) is right predictions of positive label, TN (True negative) is right predictions of negative label, FP (False positive) is wrong predictions of positive label, and FN (False negative) is wrong predictions of negative label.[24] The vital performance metrics are as follows:

a) Sensitivity (SENS)

The SENS is a measurement of the number of right positive predictions divided by the total positive's instances. It is sometimes referred to as the recall or true positive rate. The highest sensitivity is 1.0, and the lowest is 0.0. The 'Sensitivity' is calculated as shown in Eq. (1):

$$\text{SENS} = \frac{TP}{TP+FN} \quad \text{... (1)}$$

b) Specificity (SPEC)

The SPEC is measured as the number of correct negative predictions divided by the total number of negatives instances. It is sometimes referred to as true negative rate. The 'Specificity' is calculated as shown in Eq. (2):

$$\text{SPEC} = \frac{TN}{FP+TN} \quad \text{... (2)}$$

c) Precision (PREC)

The PRECis predicted as positive. The 'Precision' is calculated as shown in Eq. (3):

$$\text{PREC} = \frac{TP}{TP+FP} \quad \text{... (3)}$$

d) Negative Predictive Value (NPV)

The NPV is the probability that the prediction is correct if the prediction is negative. The 'Negative Predictive Value'is calculated as shown in Eq. (4):

$$NPV = \frac{TN}{TN+FN} \quad \text{... (4)}$$

e) False Positive Rate (FPR)

The FPR is measure as the total wrong positive predictions divided by the total negatives instance. The 'False Positive Rate' is calculated as shown in Eq. (5):

$$FPR = \frac{FP}{FP+TN} \quad \text{... (5)}$$

f) False Discovery Rate (FDR)

The FDR is defined as the predicted percentage of false positives across all proclaimed significant results.[21] The 'False Discovery Rate' is calculated as shown in Eq. (6):

$$FDR = \frac{FP}{FP+TP} \quad \text{... (6)}$$

g) False Negative Rate (FNR)

The FNR is also known as the miss rate, is the likelihood that the test will miss a real positive. The purpose of a classification model is not to bound the false positive rate. The 'False Negative Rate' is calculated as shown in Eq. (7) as follows:

$$FNR = \frac{FN}{FN+TP} \quad \text{... (7)}$$

h) Accuracy (ACC)

The ACC[22] refers to how close a measurement is to its true value. This metric parameter is used when the True negatives and True Positives are more vital.[22] The 'Accuracy' is calculated as shown in Eq. (8):

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad \ldots (8)$$

i) F1 Score (F1)

The highest value of the F1score indicates perfect recall and precision, and the lowest F1 score value is 0, which indicates that either precision or recall value is zero, this metric parameter is used when the False Positives and False Negatives are important.[23–25] The 'F1 Score' is calculated as shown in Eq. (9):

$$F1 = \frac{2TP}{2TP+FP+FN} \quad \ldots (9)$$

j) Matthews Correlation Coefficient (MCC)

The MCC explains how modifying the value of a variable affects the value of another and returns value in the range of −1 and 1. The value 1 denotes a perfect prediction, 0 denotes an inability to return any reliable information, and −1 denotes total inconsistency among both observations and prediction.[26] The 'Matthews Correlation Coefficient' is calculated as shown in Eq.:

$$MCC = \frac{(TP \times TN - FP \times FN)}{(sqrt(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)))} \quad \ldots (10)$$

Rank Determination among DMLs

The SS-DL models might be more accurate We carried out a number of tests using a medical image dataset to look into and assess the performance of our proposed method. The standard DML, such as MobileNet, ResNet18, VGG19, and VGG16, Pre-Act Resnet18 are trained and assessed from the ground up. The detail performance of standard DLM rank evaluation is presented in Table 3. The training accuracy of the standard models is shown in Fig. 5 and that of the test accuracy of standard DML in Fig. 6.

Re-train the Models using both Label and Pseudo Dataset

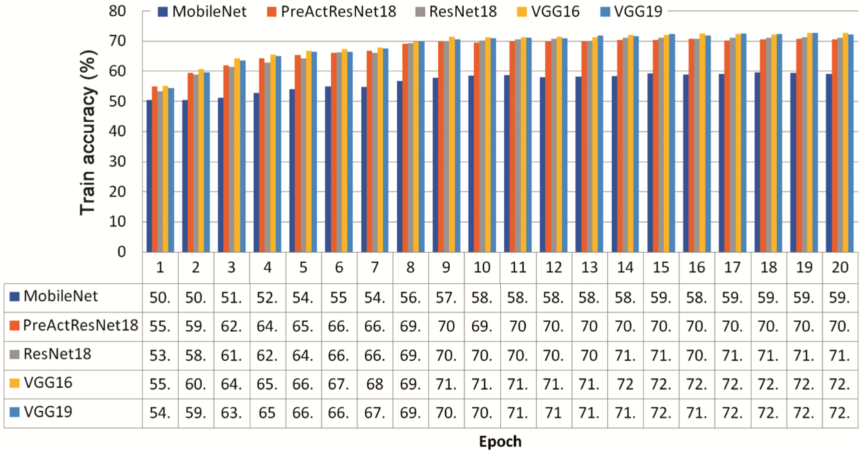
In this stage, the standard DLM are retrained with the combined MURA-SU and pseudo dataset generated by each standard DLM and our suggested model. In the next stage, validation of these trained models will be performed on the HATA-SU unseen dataset.

Validation of Propose Model on Unseen HATA-SU Dataset

The performance of the proposed model is evaluated and validated in the last step of the experiment using an unpublished dataset called HATA-SU. To evaluate the performance, the validation procedure is carried out in numerous steps. The results are depicted in Table 4. On the HATA-SU unseen dataset, we evaluate the conventional DLM's performance in the first section without using any semi-supervised learning techniques. The average measure for standard DLM such as 27.784% ACC, 7.874% SPEC, 53.068% SENS, 30.626% PREC, −44.932 % MCC, 69.374%

Table 3 — Detail performance of standard DLM rank evaluation

Models	Top 1 Train Accuracy	Top 1 Test Accuracy	Training Time for 20 Epoch	Rank
MobileNet	59.65	64.43	41 Minutes 52 Seconds	1
PreActResNet18	70.78	71.13	127 Minutes 51 Seconds	5
ResNet18	71.29	72.68	121 Minutes 39 Seconds	4
VGG16	72.88	71.64	70 Minutes 5 Seconds	2
VGG19	72.76	72.16	81 Minutes 9 Seconds	3
Threshold value (Average)	69.47	70.40	88 minutes 31 second	—



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
■ MobileNet	50.	50.	51.	52.	54.	55	54.	56.	57.	58.	58.	58.	58.	58.	59.	58.	59.	59.	59.	59.
■ PreActResNet18	55.	59.	62.	64.	65.	66.	66.	69.	70	69.	70	70.	70	70.	70.	70.	70.	70.	70.	70.
■ ResNet18	53.	58.	61.	62.	64.	66.	66.	69.	70.	70.	70.	70.	70	71.	71.	70.	71.	71.	71.	71.
■ VGG16	55.	60.	64.	65.	66.	67.	68	69.	71.	71.	71.	71.	71.	72	72.	72.	72.	72.	72.	72.
■ VGG19	54.	59.	63.	65	66.	66.	67.	69.	70.	70.	71.	71	71.	71.	72.	71.	72.	72.	72.	72.

Fig. 5 — Standard DLM train accuracy

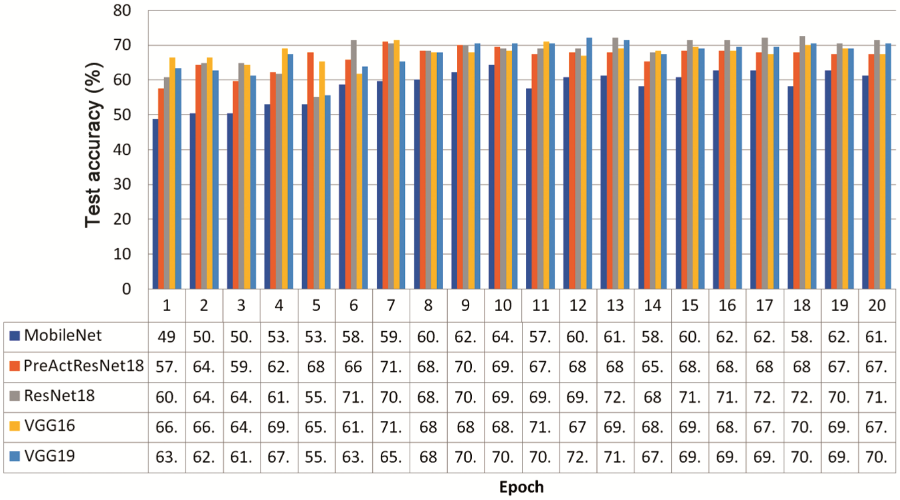


Fig. 6 — Test accuracy of standard DLM

Table 4 — Suggested model result evaluation different standard DLM

Measure		Non-semi-supervised approach	Pseudo dataset generated through					
			MobileNet	PreActResNet18	ResNet18	VGG16	VGG19	Propose model
Accuracy	MobileNet	36.27	64.61	81.94	79.88	83.85	59.32	91.7
	PreActResNet18	19.97	74.45	92.66	92.22	93.39	67.69	94.83
	ResNet18	27.17	68.58	88.11	92.95	91.19	68.72	96.89
	VGG16	21.59	79.15	94.27	92.07	91.78	75.48	95.57
	VGG19	33.92	89.53	93.69	89.72	92.8	75.33	94.89
	Average	27.784	75.264	90.134	89.368	90.602	69.308	94.77
Specificity	MobileNet	11.29	83.73	91.34	79.79	93.7	97.11	97.59
	PreActResNet18	10.5	95.8	96.33	90.29	97.9	97.11	96.85
	ResNet18	14.17	97.64	97.11	93.18	97.64	96.85	97.69
	VGG16	0.79	96.85	96.85	91.34	98.16	98.69	97.85
	VGG19	2.62	95.8	96.06	88.71	96.85	98.95	99.1
	Average	7.874	93.964	95.538	88.662	96	97.742	97.82
Sensitivity	MobileNet	68	40.33	70	80	71.33	11.33	69.33
	PreActResNet18	32	47.33	88	94.67	87.67	30.33	91
	ResNet18	43.67	31.67	76.67	92.67	83	33	93.33
	VGG16	48	56.67	91	93	83.67	46	93.67
	VGG19	73.67	79.13	90.67	91	87.67	45.33	91.33
	Average	53.068	51.026	83.268	90.268	82.668	33.198	87.732
Positive predictive value (Precision)	MobileNet	37.64	66.12	86.42	75.71	89.92	75.56	92.95
	PreActResNet18	21.97	89.87	94.96	88.47	97.05	89.22	96.74
	ResNet18	28.6	91.35	95.44	91.45	96.51	89.19	98.88
	VGG16	27.59	93.41	95.79	89.42	97.29	96.5	95.95
	VGG19	37.33	91.92	94.77	86.39	95.64	97.14	98.01
	Average	30.626	86.534	93.476	86.288	95.282	89.522	96.506
Matthews correlation coefficient	MobileNet	−25.51	26.95	63.56	59.51	67.71	16.88	72.27

(*contd.*)

Table 4 — Suggested model result evaluation different standard DLM (*contd.*)

Measure		Non-semi-supervised approach	Pseudo dataset generated through					
			MobileNet	PreActRes Net18	ResNet18	VGG16	VGG19	Propose model
	ResNet18	−44.6	40.44	76.6	85.73	82.52	40.12	93.74
	VGG16	−60.1	60.04	88.41	84.03	83.74	54.47	89.99
	VGG19	−34.92	77.57	87.2	79.35	85.51	54.4	92.68
	Average	−44.932	51.146	80.188	78.622	81.254	40.81	87.444
False discovery rate	MobileNet	62.36	33.88	13.58	24.29	10.08	24.44	5.05
	PreActResNet18	78.03	10.13	5.04	11.53	2.95	10.78	3.96
	ResNet18	71.4	8.65	4.56	6.82	3.49	10.81	0.97
	VGG16	72.41	6.59	4.21	8.66	2.71	3.5	4.18
	VGG19	62.67	8.08	5.23	11.29	4.36	2.86	2.98
	Average	69.374	13.466	6.524	11.338	4.7	10.478	3.428
False positive rate	MobileNet	88.71	16.27	8.66	20.21	6.3	2.89	3.71
	PreActResNet18	89.5	4.2	3.67	9.71	2.1	2.89	3.05
	ResNet18	85.83	2.36	2.89	8.55	2.36	3.15	1.01
	VGG16	99.21	3.15	3.15	10.58	1.84	1.31	3.05
	VGG19	97.38	4.2	3.94	13.61	3.15	1.05	1.71
	Average	92.126	6.036	4.462	13.712	3.15	2.258	2.506
F1 Score	MobileNet	48.46	50.1	77.35	77.8	79.55	19.71	79.45
	PreActResNet18	26.05	62.01	91.35	91.47	92.12	45.27	93.78
	ResNet18	34.56	47.03	85.03	92.05	89.25	48.18	96.19
	VGG16	35.04	70.54	93.33	91.18	89.96	62.3	94.7
	VGG19	49.55	85.05	92.67	88.64	91.48	61.82	96.24
	Average	38.732	62.946	87.946	88.228	88.472	47.456	92.072
Negative predictive value	MobileNet	30.94	64.06	79.45	83.52	80.59	58.18	79.97
	PreActResNet18	16.39	69.79	91.07	95.56	90.98	63.9	93.48
	ResNet18	24.22	64.47	84.09	94.16	87.94	64.74	95.24
	VGG16	1.89	73.95	93.18	94.31	88.42	69.89	94.65
	VGG19	11.24	88.38	92.89	92.6	90.89	69.69	95.91
	Average	16.936	72.13	88.136	92.03	87.764	65.28	91.85
False negative rate	MobileNet	32	59.67	30	20	28.67	88.67	33.67
	PreActResNet18	68	52.67	12	5.33	12.33	69.67	9
	ResNet18	56.33	68.33	23.33	7.33	17	67	6.67
	VGG16	52	43.33	9	7	16.33	54	7.33
	VGG19	26.33	20.87	9.33	9	12.33	54.67	5.67
	Average	46.932	48.974	16.732	9.732	17.332	66.802	12.468

FDR, 92.126% FPR, 38.732% f1, 16.936% NPV, and 46.932% FNR.

The results of the traditional DLM on the HATA-SU unobserved data is evaluated in the second section when a pseudo data is label using MobileNet. The average measure for standard DLM such as 75.264% ACC, 93.964% SPEC, 51.026 SENS, 86.534% PREC, 51.146% MCC, 13.466% FDR, 6.036% FPR, 62.946% f1, 72.13% NPV, and 48.974% FNR. The results of the traditional DLM on the HATA-SU unobserved data is evaluated in the third section when a pseudo data is label using PreActResNet18. The average measure for standard DLM such as 90.134% ACC, 95.538% SPEC, 83.268% SENS, 93.476% PREC, 80.188% MCC, 6.524% FDR, 4.462% FPR, 87.946% f1, 88.136% NPV, and 16.732% FNR. The results of the traditional DLM on the HATA-SU unobserved data is evaluated in the fourth section

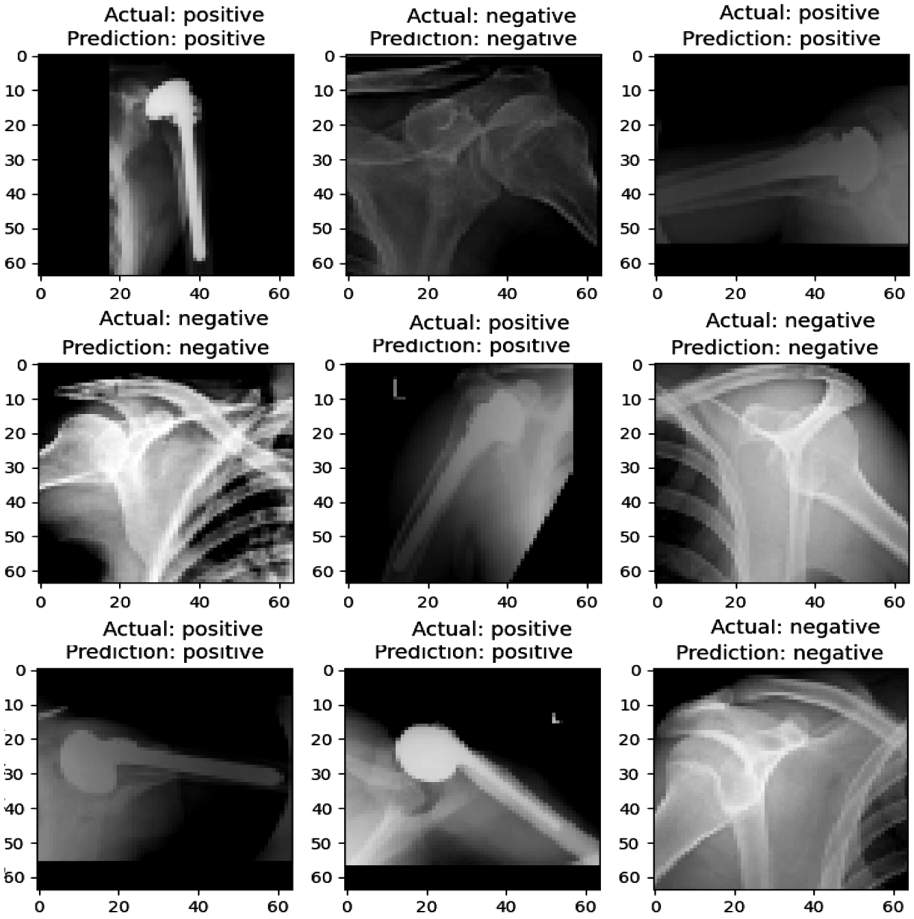


Fig. 7 — Correct and incorrect prediction label through suggested model with Actual vs Prediction as (a) p & p, (b) n & n, (c) p & p, (d) n & n, (e) p & p, (f) n & n, (g) p & p, (h) p & p, (i) n & n. Here p = positive and n = negative.

when a pseudo data is label using ResNet18. The average measure for standard DLM such as 89.368% ACC, 88.662% SPEC, 90.268% SENS, 86.288% PREC, 78.622% MCC, 13.712% FDR, 11.338% FPR, 88.228% f1, 92.03% NPV, and 9.732% FNR. The results of the traditional DLM on the HATA-SU unobserved data is evaluated in the fifth section when a pseudo data is label using VGG16. The average measure for standard DLM such as 90.602% ACC, 96.85% SPEC, 82.668% SENS, 95.282% PREC, 81.254% MCC, 4.718% FDR, 3.15% FPR, 88.472% f1, 87.764% NPV, and 17.332% FNR. The results of the traditional DLM on the HATA-SU unobserved data is evaluated in the sixth section when a pseudo data is label using VGG19. The average measure for standard DLM such as 69.308% ACC, 94.77% SPEC, 97.742% SENS, 33.198% PREC, 89.522% MCC, 40.81%FDR, 10.478% FPR, 47.456% f1, 65.28% NPV, and 65.28% FNR.The results of the traditional DLM on the HATA-SU unobserved data is evaluated in the seventh section when a pseudo data is label using our suggested model. The average measure for standard DLM are 94.77% ACC, 97.82% SPEC, 87.732% SENS, 96.506% PREC, 87.444% MCC, 3.428% FDR, 2.506% FPR, 92.072% f1, 91.85% NPV, and 12.468% FNR. Numerous illustrative cases along with the correct and incorrect predictions made by our suggested model for classifying shoulder bone fractures using the concealed HATA-SU dataset is shown in Fig. 7 from where it is evident that the model's classification performance might be enhanced. In contrast to more intricate model designs, we mostly used the fundamental DLM in our work. since we're looking at methods to use unlabeled data efficiently to aid the medical sector.

Overall, the recommended models performed well, with the semi-supervised strategy not being the least effective when compared to the others. The results of the traditional DLM on the HATA-SU unobserved

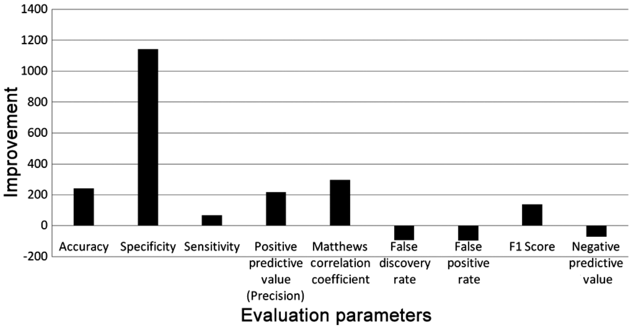


Fig. 8 — Performance enhanced from no semi-supervised approach and Suggested model (RTS-SS-DL)

data is evaluated when a pseudo data is label using our suggested model (Table 4). The average measure for standard DLM are 94.77% ACC, 97.82% SPEC, 87.732% SENS, 96.506% PREC, 87.444% MCC, 3.428% FDR, 2.506% FPR, 92.072% f1, 91.85% NPV, and 12.468% FNR. The suggested model achieves an increase of 241.09% ACC, 1142.31% SPEC, 65.31% SENS, 215.11% PREC, 137.71% f1, 294% MCC, 241.09% NPV, and decreases of −95.05% FDR,−97.27% FPR, and −73.43% FNR (Fig. 8).

Limitation

The use of ensemble methods is an effective approach to improve the effectiveness and robustness of DLM. Nevertheless, there are restrictions that must be taken into account while utilizing these techniques. The following are few restrictions associated with proposed methods: The complexity of the suggestive approach can be significantly increased due to the need to train multiple models, which can also extend the time needed to evaluate the models. These techniques can be prone to over fitting if structure is complex or if there is a shortage of training data, leading to poor result on new data. The explainability of proposed methods is often limited as they are typically considered "black box" models, making it difficult to explain the outcomes. The overall performance of proposed model is heavily dependent on the performance of DML. Poor performance by one or more models can negatively impact the overall performance of the proposed model.

Conclusions

In this research, we offer RTS-SSDL, Rank-based Two-Stage Semi-Supervised Deep Learning framework for X-ray data analysis that makes use of Active Learning (AL) to unlock the potential of unlabeled data. We suggest adaptive pseudo-labeling and instructive active annotation in the algorithm, which make use of the unlabeled medical pictures and create a closed-loop structure to enhance the performance of the SSL model for medical images. The proposed model outperforms the other models when both labeled and pseudo datasets are used. While accuracy is a useful performance statistic, it may not always be the best when the dataset's target variable classes are imbalanced. Therefore, the study also considers other metrics, such as FPR, FNR, FDR, SENS, SPEC, PREC, and MCC, which indicate an effective and efficient model. The proposed RTS-SSDL models have the potential to reduce labeling efforts while maintaining excellent model performance. According to the experimental findings, RTS-SSDL performs noticeably better on medical image classification tasks than other approaches. In terms of future directions, the presented RTS-SSDL model can be extended to different medical imaging datasets to evaluate its effectiveness in other domains. Additionally, the model's architecture can be further optimized to improve its performance and reduce the computational requirements.

The study only evaluates the model's performance on a binary classification task, so future work could explore its performance on multi-class classification tasks. Furthermore, exploring the use of transfer learning to enhance the model's ability to generalize to new datasets would be another interesting direction. Overall, the proposed RTS-SSDL model shows promise for the semi-supervised classification of medical imaging data, and there is significant potential for further development and optimization in future research.

References

1 Agrawala A, Learning with a probabilistic teacher, *IEEE Trans Inf Theory*, **16(4)** (1970) 373–379, https://doi.org/10.1109/TIT.1970.1054472.

2 Fralick S, Learning to recognize patterns without a teacher,*IEEE Trans Inf Theory,***13(1)** (1967) 57–64, https://doi.org/10.1109/TIT.1967.1053952.

3 Scudder H, Probability of error of some adaptive pattern-recognition machines, *IEEE Trans Inf Theory*,**11(3)** (1965) 363–371,https://doi.org/10.1109/TIT.1965.1053799.

4 Rajpurkar P, Irvin J, Bagul A, Ding D, Duan T, Mehta H, Yang B, Zhu K, Laird D, Ball R L & Langlotz C, MURA: Large dataset for abnormality detection in musculoskeletal radiographs,*arXiv Prepr arXiv171206957*, (2017).

5 Narayan V & Daniel A K, CHHP: coverage optimization and hole healing protocol using sleep and wake-up concept for wireless sensor network, *Int J Syst Assur Eng Manag*, **13(1)** (2022) 546–556.

6 Narayan V & Daniel A K, IOT based sensor monitoring system for smart complex and shopping malls, in *Int Conf Mobile Netw Manag* (Cham: Springer International Publishing) 2021, 344–354.

7 Narayan V & Daniel A K, A novel approach for cluster head selection using trust function in WSN, *Scalable Comput Pract Exp*, **22(1)** (2021) 1–13.

8 Laine S & Aila T, Temporal ensembling for semi-supervised learning, *arXiv Prepr arXiv161002242*, (2016).

9 Narayan V & Daniel A K, Multi-tier cluster based smart farming using wireless sensor network, in *2020 5th Int Conf Comput, Commun Secur* (IEEE) 2020, 1–5, https://doi.org/ 10.1109/ICCCS49678.2020.9277072.

10 Awasthi S, Srivastava A P, Srivastava S & Narayan V, A Comparative study of various CAPTCHA methods for securing web pages, in *2019 Int Conf Automat Comput Technol Manag* (IEEE) 2019, 217–223, https://doi.org/10.1109/ICACTM.2019.8776832.

11 Narayan V & Daniel A K, Design consideration and issues in wireless sensor network deployment, *Invertis J Sci & Technol*, (2020) 101–109.

12 Tarvainen A & Valpola H, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, *arXiv Prepr arXiv170301780*, (2017).

13 Miyato T, Maeda S, Koyama M & Ishii S, Virtual adversarial training: a regularization method for supervised and semi-supervised learning, *IEEE Trans Pattern Anal Mach Intell*, **41(8)** (2018) 1979–1993, https://doi.org/10.1109/TPAMI.2018.2858821.

14 Berthelot D, Carlini N, Goodfellow I, Papernot N, Oliver A & Raffel C, Mixmatch: A holistic approach to semi-supervised learning, *arXiv Prepr arXiv190502249*, Published online (2019).

15 Howard A G, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M & Adam H, Mobilenets: Efficient convolutional neural networks for mobile vision applications, *arXiv Prepr arXiv170404861*, (2017).

16 He K, Zhang X, Ren S & Sun J, Identity mappings in deep residual networks, in *Computer Vision–ECCV 2016: 14th Euro Conf, Amsterdam, The Netherlands, October 11–14, 2016, Proc, Part IV 14* (Springer International Publishing 2016, 630–645, http://arxiv.org/abs/1603.05027.

17 He K, Zhang X, Ren S & Sun J, Deep residual learning for image recognition, *CoRR,* (2015) abs/1512.0, http://arxiv.org/abs/1512.03385.

18 Zhang X, Zou J, He K & Sun J, Accelerating very deep convolutional networks for classification and detection, *IEEE Trans Pattern Anal Mach Intell*, **38(10)** (2015) 1943–1955.

19 Simonyan K & Zisserman A, Very deep convolutional networks for large-scale image recognition, *arXiv Prepr arXiv14091556*,(2014).

20 Setiawan F, Yahya B N & Lee S L, Deep activity recognition on imaging sensor data, *Electron Lett,* **55(17)** (2019) 28–931.

21 Benjamini Y, Discovering the false discovery rate, *J R Stat Soc Ser B statistical Methodol*, **72(4)** (2010) 405–416.

22 Mall P K, Singh P K & Yadav D, GLCM based feature extraction and medical X-RAY image classification using machine learning techniques, in *2019 IEEE Conf Info Commun Technol*(IEEE) 2019, 1–6.

23 Narayan V & Daniel A K, CHOP: Maximum coverage optimization and resolve hole healing problem using sleep and wake-up technique for WSN, *Adv Distrib Comput Artif Intell J*, **11(2)** (2022) 159–178.

24 Narayan V & Daniel A K, RBCHS: Region-based cluster head selection protocol in wireless sensor network, in *Proc Integrat Intell Enable Netw Comput* (Springer) 2021, 863–869.

25 Srivastava S & Sharma S, Analysis of cyber related issues by implementing data mining algorithm, in *2019 9th Int Conf Cloud Comput Data Sci Eng* (IEEE) 2019, 606–610, https://doi.org/10.1109/CONFLUENCE.2019.8776980.

26 Chicco D & Jurman G, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, *BMC Genomics*, **21(1)** (2020) 1–13.