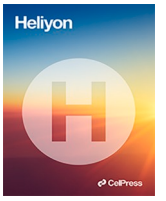
Heliyon 10 (2024) e38910

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

MIDC: Medical image dataset cleaning framework based on deep learning

Sanli Yi [a,b,*], Ziyan Chen [a,b]

[a] School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, China
[b] Key Laboratory of Computer Technology Application of Yunnan Province, Kunming, 650500, China

ARTICLE INFO

Keywords:
Deep learning
Data cleaning
Public medical datasets
Mislabeled data
Classification accuracy

ABSTRACT

Deep learning technology is widely used in the field of medical imaging. Among them, Convolutional Neural Networks (CNNs) are the most widely used, and the quality of the dataset is crucial for the training of CNN diagnostic models, as mislabeled data can easily affect the accuracy of the diagnostic models. However, due to medical specialization, it is difficult for non-professional physicians to judge mislabeled medical image data. In this paper, we proposed a new framework named medical image dataset cleaning (MIDC), whose main contribution is to improve the quality of public datasets by automatically cleaning up mislabeled data. The main innovations of MIDC are: firstly, the framework innovatively utilizes multiple public datasets of the same disease, relying on different CNNs to automatically recognize images and remove mislabeled data to complete the data cleaning process. This process does not rely on annotations from professional physicians and does not require additional datasets with more reliable labels; Secondly, a novel grading rule is designed to divide the datasets into high-accuracy datasets and low-accuracy datasets, based on which the data cleaning process can be performed; Thirdly, a novel data cleaning module based on CNN is designed to identify and clean low-accuracy datasets by using high-accuracy datasets. In the experiments, the validity of the proposed framework was verified by using four kinds of datasets diabetic retinal, viral pneumonia, breast tumor, and skin cancer, with results showing an increase in the average diagnostic accuracy from 71.18 % to 85.13 %, 82.50 %to 93.79 %, 85.59 %to 93.45 %, and 84.55 %to 94.21 %, respectively. The proposed data cleaning framework MIDC could better help physicians diagnose diseases based on the dataset with mislabeled data.

1. Introduction

With the continuous development of deep learning technology, its application in many fields such as image processing [1], natural language processing [2], industrial [3], fault detection [4], multimedia [5], and so on has achieved remarkable results. In the field of medical image processing [6], convolutional neural networks are trained on sample data with doctor diagnostic labels to obtain diagnostic models. Thus, the quality of the dataset is crucial for the model [7], including data amount and accurate labeling [8]. The large data amount provides sufficient feature information [9] and an accurate label ensures the model learns features more correctly

* Corresponding author. School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, China.
E-mail address: 152514845@qq.com (S. Yi).

https://doi.org/10.1016/j.heliyon.2024.e38910
Received 25 February 2024; Received in revised form 17 September 2024; Accepted 2 October 2024
Available online 3 October 2024

[10], whereas the presence of incorrect labels in the data leads to an unreliable model [11], which in turn affects the diagnostic results.

The concept of data cleaning was first introduced by Galhardas in 2000 to address problems and errors in data [12]. At first, this technique is used in the text data. Such as: Jason et al. [13] utilized Pandas for missing data processing and outlier recognition, improving data quality through feature encoding and standardization techniques to enhance the performance of machine learning models. Konstantinos et al. [14] proposed a multi-level electronic medical record (EMR) data cleaning method, which processes missing values in EMR through deletion and interpolation, and processes outliers in EMR through cluster analysis. All these methods are aimed at text data, while not suitable for image data. In recent years, researchers have gradually explored the use of deep learning for image data cleaning [15]. Currently, the research on data cleaning focuses on two directions: One is to deal with the data belonging to minority classes or with low-resolution of datasets, which are labeled correctly but affect the model training results due to its small amount or bad quality [16]. Studies in this direction include: In 2020, Zhang et al. [17] proposed the ImageDC data cleaning method based on the Twitter dataset. This method first calculates the average and standard deviation of the number of images in each category, and then sets category thresholds to identify and clean minority class data with quantities below the threshold. Then calculate the self-recognition rate of each category and set a recognition threshold to identify and clean low-resolution data below the threshold. The accuracy of 542 categories was ultimately improved from 64.17 % to 68.88 %. Liu et al. [18] used threshold method for data cleaning and used pre-training Xception to remove a few categories of data from the South China Sea marine fish dataset, improving the accuracy of marine fish classification to 75.27 %. However, the cleaning effect of the above methods depends on the threshold setting, and different datasets require different thresholds. The selection of these parameters has a significant impact on the cleaning results. Meanwhile, the above methods are only applicable to a few categories of images and are not suitable for medical image datasets. Another method is to clear the mislabeled data of the dataset, which provides inaccurate feature information and leads to incorrect classification results when training the network on the dataset [19]. For example, Curtis et al. [20] proposed confidence learning, which iteratively cleans up mislabeled data by using a given dataset with accurate labels. But this data cleaning method is difficult to implement because it requires additional datasets with correct labels, which are often difficult to obtain. Based on cluster analysis, Li et al. [21] used VGG-NIN neural network to identify the outlier data as mislabeled data and cleared them. Outliers are data points that deviate significantly from other data, usually indicating label errors. By identifying these outliers and treating them as potential instances of mislabeling, the early esophageal cancer dataset classification accuracy is improved to 85 %. This method can clear some mislabeled data, but as correctly labeled data may also be identified as outlier, it may miss useful data.

In the field of natural images, due to the abundant amount of data, the problem of datasets has not yet become prominent [22]. However, for the medical images, data cleaning is particularly important for the following reasons: (1) Since it's difficult to obtain enough medical data with diagnostic information [23], the size of medical datasets is usually small [24], and even a small number of mislabeled data in it can introduce a large impact. (2) The process of labeling medical data is highly specialized, i.e., doctors with different levels of experience and sophistication can label a piece of data differently [25], whereas non-medical researchers are unable to identify the correct one. (3) In the field of medical diagnosis, there are usually multiple public datasets for a given disease [26]. For example, public datasets like APTOS [27], Messidor-2 [28], Eyepac [29], and STARE [30] are available for diabetic retinal diseases, Chest X-Ray Images [31], COVID-19 [32] are available for viral pneumonia, and us-data [33], BUSI [34], BreakHis [35] are available for breast tumors. Although these datasets are open, their performance on the same network are different due to the different accuracy of their data labels. For example, in the researches of Yue et al. [36] and Lahmar et al. [37], the accuracy of APTOS and Messidor-2 diagnosis on ResNet50 differs by 27.44 %. In the researches of Li et al. [38] and Lahmar et al. [39], the accuracy of Chest X-Ray Images and COVID-19 diagnosis on ResNet50 differs by 5.08 %. In the researches of Masud et al. [40] and Moon et al. [41], the accuracy of us-data and BUSI diagnosis on ResNet50 differs by 13.55 %. This issue will affect the use of these datasets because researchers prefer to

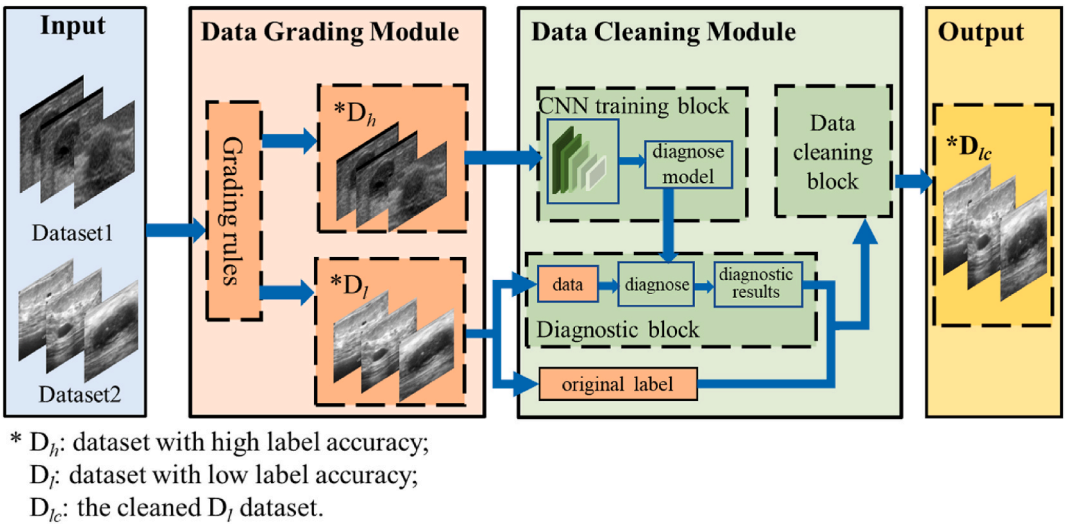


Fig. 1. MIDC functional structure diagram.

use the better one rather than the poorer one [42]. While for poor dataset that contain majority of valid data, it makes more sense to improve its quality by cleaning the bad-data rather than giving up to use it [43].

Aimed at above problems, a medical data cleaning framework MIDC is proposed to improve the quality of public medical datasets by cleaning mislabeled data, and by which more effective medical diagnose model can be given. The main contributions of the framework are as follows: (1) The framework innovatively uses multiple public datasets of the same disease to perform the data cleaning process, through which the datasets can be easily accessed and do not need extra datasets with more reliable labels. (2) We propose a novel grading rule to divide the datasets into high-accuracy datasets and low-accuracy datasets. And based on these results, data cleaning process can be performed. (3) A novel data cleaning module is designed to clean low-accuracy datasets by using high-accuracy datasets.

The rest of this paper is described as follows: Section 2 provides a detailed description of the proposed framework and materials; Section 3 presents the experiments and results of the validation framework; Section 4 analyzes the results of all the experiments; Section 5 summarizes the significance of this work and future improvement directions.

2. Framework and materials

The medical image dataset cleaning framework MIDC proposed in this paper is built on the basis of multiple public datasets that represent the same kind of diseases but have varying labeling qualities. The functional structure diagram of MIDC is shown in Fig. 1, which mainly includes the input part, data grading module, data cleaning module, and output part. The specific description is as follows: Firstly, we input two datasets of the same disease, Dataset 1 and Dataset 2; Secondly, based on the grading rules, the data grading module classifies the two datasets into dataset D_h with higher label accuracy and dataset D_l with lower label accuracy. Then D_h and D_l are put into the data cleaning module separately; Thirdly, the data cleaning module which are composed of CNN training block, diagnostic block and data cleaning block, cleans the D_l by using D_h to obtain a cleaned dataset D_{lc} with a higher labeling accuracy; Finally, D_{lc} is output. The pseudocode of the MIDC framework which demonstrates the design information is shown in Appendix A. And the flowchart of the MIDC framework is shown in Appendix B.

2.1. Input datasets

The input dataset of the MIDC framework should meet the following guidelines: (1) Since the framework uses high-accuracy datasets to clean low-accuracy datasets to better diagnose a specific disease, it must rely on at least two or more publicly available datasets. (2) The dataset should satisfy that there is a sufficient amount of data in the dataset and sufficient literature citing the dataset. The former ensures that the neural network can effectively train a diagnose model, and the latter serves as a criterion for evaluating the quality of the dataset.

Currently, with the wide application of deep learning in medical diagnosis, public datasets for many diseases meet the above requirements. In this paper, four kinds of diseases are selected for the study, including diabetic retinal, viral pneumonia, breast tumors and skin cancer. The category and number statistics for each dataset are shown in Table 1, and example images for each dataset are displayed in Fig. 2.

A. Diabetic retinal

The dataset in this category consists of retinal scan images used to detect diabetic retinopathy. Based on the physician's diagnosis,

Table 1
The number of each category in different datasets.

Disease type	Dataset	Category	Number	Total
Diabetic retinal	APTOS	normal	1805	3662
		diabetes	1857	
	Messidor-2	normal	1018	1786
		diabetes	768	
Viral pneumonia	Chest X-Ray Images	normal	1585	5860
		pneumonia	4275	
	COVID-19	normal	1341	2905
		pneumonia	1564	
Breast tumors	us-data	benign	100	250
		malignant	150	
	BUSI	benign	437	647
		malignant	210	
Skin cancer	ISIC 2020	benign	584	1168
		malignant	584	
	MED NODE	benign	400	680
		malignant	280	

*The examples of above datasets are show in Fig. 2: Fig. 2(a) for APTOS; Fig. 2(b) for Messidor-2; Fig. 2(c) for Chest X-Ray Images; Fig. 2(d) for COVID-19; Fig. 2(e) for us-data; Fig. 2(f) for BUSI; Fig. 2(g) for ISIC2020; Fig. 2(h) for MED NODE.

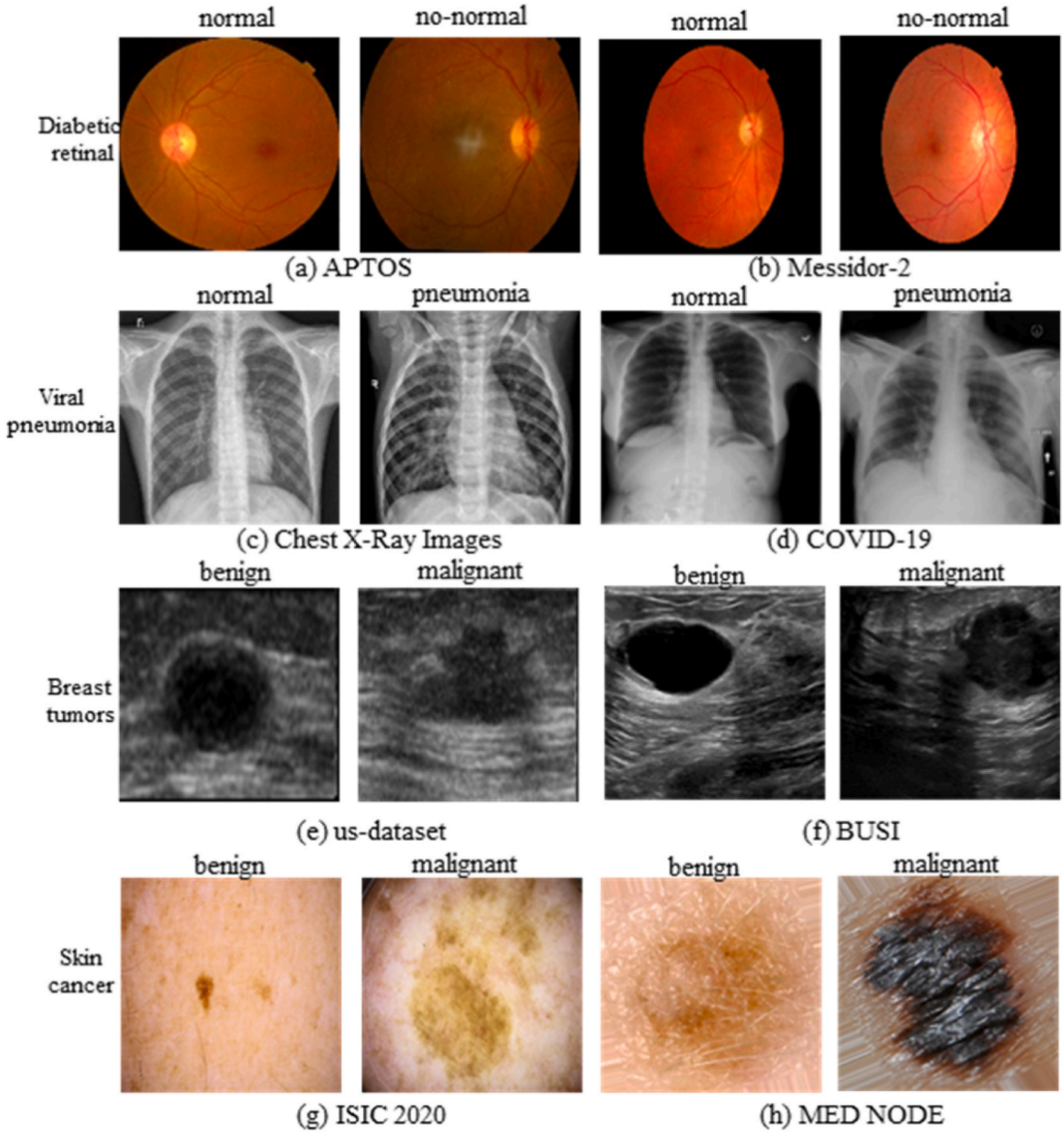


Fig. 2. Examples of images from the datasets belong to the four diseases.

these images are labeled as grade 0–4, reflecting the severity of diabetic retinopathy. The datasets used in this paper are: (1) APTOS [27], which is provided by Aravind Eye Hospital, India, containing 3662 images; (2) Messidor-2 [28], which is provided by ADCIS, containing 1786 images.

B. Viral pneumonia

The dataset in this category consists of chest x-ray images used to detect pneumonia. Based on the physician's diagnosis, these images are labeled as normal and viral pneumonia. The datasets used in this paper are: (1) Chest X-Ray Images [31], retrospective cases from pediatric patients aged 1–5 years old from Guangzhou Women's and Children's Medical Center, containing 5860 images; (2) COVID-19 [32], provided by researchers from Qatar University and University of Dhaka, Bangladesh, containing 2905 images.

C. Breast tumors

The dataset in this category consists of ultrasound images of breast tumors used to detect breast cancer. The images are labeled as benign and malignant based on the physician's diagnosis. The datasets used in this paper are: (1) us-data [33], from Mendeley, containing 250 images. (2) BUSI [34], from Al-Dhabyani et al. containing 647 images.

D. Skin cancer

This category of dataset includes dermatoscopic images used for detecting skin cancer. According to the doctor's diagnosis, these images are marked as benign and malignant. The datasets used in this paper are: (1) ISIC 2020 [44], published by the International Skin Imaging Collaboration (ISIC), which includes 1168 images of benign and malignant skin lesions. (2) MED NODE [45], the images in this dataset are from the Department of Dermatology at Groningen University Medical Center in the Netherlands, with a total of 70 malignant and 100 benign dermatoscopic images. Due to its small data size, we expand the dataset by rotating the images by 15°, 45°, and 90°. The expanded datasets are shown in Table 1.

Fig. 2 represents example images for each category in each dataset: (1) For diabetic retinal, Fig. 2(a) and (b) is example of APTOS and Messidor-2 respectively, in which the left is normal image and the right is abnormal image; (2) For viral pneumonia, Fig. 2(c) and (d) is example of Chest X-Ray and COVID-19, in which the left is normal image and the right is viral pneumonia image; (3) For breast tumors, Fig. 2(e) and (f) is example of us-data and BUSI, in which the left is benign breast tumor image and the right is malignant breast tumor image; (4) For skin cancer, Fig. 2(g) and (h) is example of ISIC 2020 and MED NODE, in which the left is benign image and the right is malignant image.

2.2. Data grading module

In the framework, two different public datasets are input to the data grading module, which divides them into high-accuracy dataset D_h and low-accuracy dataset D_l. The rules of data grading are based on two guidelines: (1) Citation guideline, according to which we first collect literature based on these datasets in recent years and then distinguish the quality of these datasets by comparing their different performances; (2) Experiment guideline, according to which we first use these datasets to train the common neural networks and then distinguish the quality of these datasets by comparing their experimental results.

2.3. Data cleaning module

For the obtained two levels of dataset D_h and D_l, the data cleaning module uses the high-accuracy dataset D_h as the baseline to clean the low-accuracy dataset D_l, thereby improving the quality of its data labels. The structure of the data cleaning module is shown in Fig. 1: Firstly, the high-accuracy dataset D_h are put into CNN training block, which trains the CNN networks on D_h to obtain the diagnose models; Secondly, together with the low-accuracy dataset D_l, these diagnostic models are put into the diagnostic block to obtain the diagnostic results of D_l; Thirdly, these diagnostic results, along with the original label of D_l, are put into the data cleaning block, which cleans the dataset D_l based on them to obtain a dataset D_{lc} with a higher labeling accuracy. The pseudocode of the data cleaning module which demonstrate the design information is shown in Appendix C, and the detailed introduction of diagnostic block and data cleaning blocks are as follows:

A. Diagnostic block

In the diagnostic block, the diagnose model are used to diagnose D_l data to obtain the diagnostic results of D_l. In this paper, two kinds of diagnose methods are designed: multi-training and multi-network to diagnose D_l data, as shown in Fig. 3(a) and (b).

Fig. 3(a) shows the multi-training method: firstly, we train the diagnose model n times based on D_h to get n diagnose models;

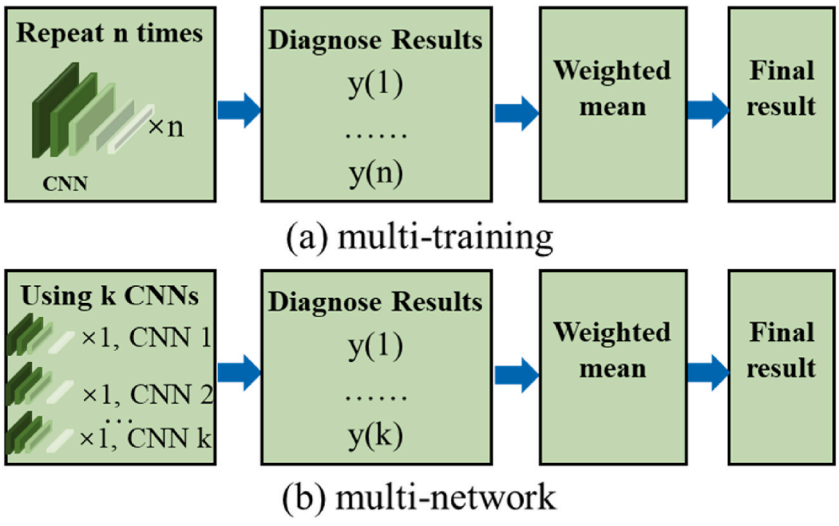


Fig. 3. Two methods of diagnose include (a) multi-training, (b) multi-network.

secondly, dataset D_l are put into these models respectively to get n diagnose results $y(i)(i = 1,2,\ldots\ldots,n)$; thirdly, based on these diagnose results, the weighted averaging result y_{mt} is calculated by equation (1). The core idea is that through repeated training, the network will gradually adjust its parameters to improve its overall generalization ability.

$$y_{mt} = \frac{\sum_{i=1}^{n} w_i y(i)}{\sum_{i=1}^{n} w_i} \tag{1}$$

Where $y(i)$ represents diagnostic result of each time, and w_i represents the weight assigned to each $y(i)$. And $y(i)$ is calculated according to equation (2) [46], which takes a binary task as an example:

$$y(i) = p0 \times c0 + p1 \times c1; c = 0,1 \tag{2}$$

Where c represents the classification of the prediction label. As there are two classes in such task, we can set $c = 0,1$. $c0$ represents the prediction label as class 0, $c1$ represents the prediction label as class 1, $p0$ represents the probability of $c0$, and $p1$ represents the probability of $c1$.

Fig. 3(b) shows the multi-network method: firstly, we put dataset D_h into k different networks to get k diagnose models; secondly, dataset D_l are put into these models respectively to get k diagnose results $y(i)(i = 1,2,\ldots\ldots,k)$; thirdly, based on these diagnose results, the weighted averaging result y_{mn} is calculated by equation (3). The core idea is that each network has its own unique architecture and parameter settings. This approach aims to improve the overall diagnostic performance through diversity.

$$y_{mn} = \frac{\sum_{i=1}^{k} w_i y(i)}{\sum_{i=1}^{k} w_i} \tag{3}$$

Based on y_{mt} and y_{mn}, the final diagnostic result y_{dr} are obtained by equation (4).

$$y_{dr} = \{y_{mt} \; or \; y_{mn}\} \tag{4}$$

B. Data cleaning block

In order to clean the dataset D_l to obtain the dataset D_{lc} with higher labeling accuracy, the data cleaning block performs data cleaning process as shown in Fig. 4, which is mainly consists of 3 steps: Step 1, compare, by comparing the original and diagnostic labels of each image, we can identify the image that is inconsistent with the original and diagnostic labels; Step 2, marking error data, due to the inconsistencies between the original labels and diagnostic labels, these images are considered suspicious and their original labels are considered unreliable, thus they are marked as error data; Step 3, delete error data, because these error data can affect the accuracy of diagnose model, we delete them. Then the remained data is the dataset D_{lc} with higher accuracy.

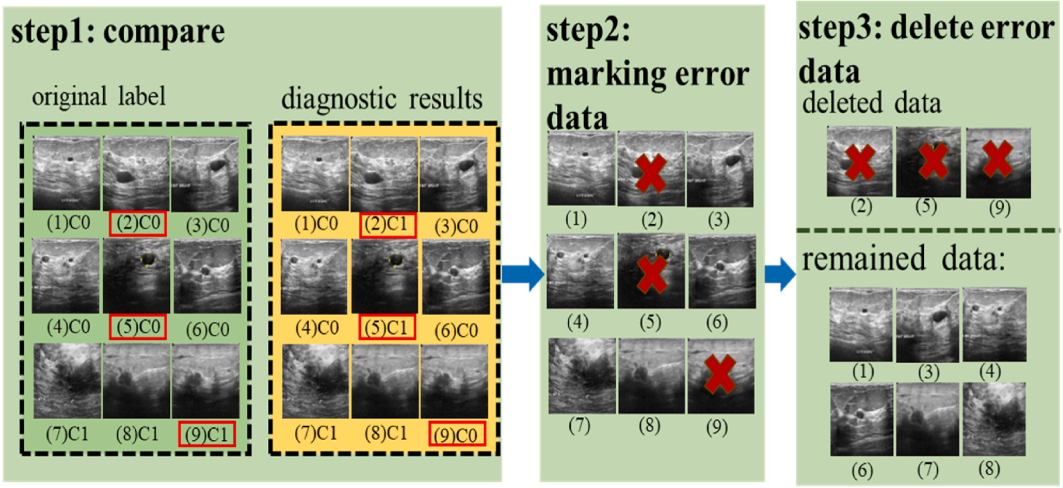


Fig. 4. Data cleaning block diagram.

3. Experiments

3.1. Experimental setup

This paper employs Pycharm as the compiler, programming language Python, experimental framework Keras and Tensorflow, and hardware environment 11th Gen Intel (R) Core (TM) i7-11700K CPU@3.60GHz The graphics card is NVIDIA GeForce RTX 3060, operating on a 64-bit Windows system. For classification experiments, we divided the entire dataset into an 80:20 training set and a testing set. Set the number of bootstrap iterations to 60. The Adam optimizer was employed for parameter updates in the context of spark_categorical_crosstropy as the loss function. Specific hyperparameters are configured with a learning rate of 0.001, 200 epochs, and a batch size of 16.

3.2. Evaluation indicators

The performance of the data cleaning framework is assessed using diverse performance indicators. For classification experiments, for better comparison, we evaluate the performance of the proposed classification network using accuracy, recall, precision, and F1 value [47]. Accuracy provides the proportion of the model's overall correct predictions, as shown in equation (5). While recall and precision measure the model's ability to identify positive classes and avoid misidentifying negative classes as positive classes, respectively, as shown in equation (6) and equation (7). The F1 value is the harmonic mean of precision and recall, as shown in equation (8). The formula for each indicator is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{8}$$

Table 2
Results of references for four kinds of disease datasets.

Disease type	Network	Datasets	
		APTOS accuracy(%)	Messidor-2accuracy(%)
Diabetic retinal	ResNet50	[a]98.27 [36],96.66 [49]	70.83 [37],78.21 [50]
	VGG16	92.91 [36]	81.59 [37],79.32 [50]
	VGG19	93.00 [49]	79.46 [37]
	InceptionV3	94.46 [36]	72.79 [37],74.09 [50]
	InceptionResNetV2	97.54 [48]	72.87 [37]
	Xception	97.81 [48]	78.77 [50]
	MobileNetV2	97.68 [48],93.17 [49]	83.81 [37]
	average	**95.72**	**77.17**
Viral pneumonia	–	**Datasets**	
		Chest X-Ray Images accuracy(%)	**COVID-19 accuracy(%)**
	VGG16	97.80 [51],98.11 [52]	83.71 [52],85.26 [39]
	ResNet18	98.47 [51],98.11 [52]	83.14 [52],88.42 [39]
	ResNet50	97.71 [38]	92.63 [39]
	DenseNet161	96.79 [38],98.97 [52]	83.52 [52]
	average	**97.99**	**86.11**
Breast tumors	–	**Datasets**	
		us-data accuracy(%)	**BUSI accuracy(%)**
	VGG16	100 [40]	82.80 [54],88.72 [41]
	ResNet18	99.50 [53],100 [40]	86.65 [41]
	ResNet50	99.00 [53],99.60 [40]	88.85 [54],86.05 [41]
	ResNet101	99.00 [53]	86.05 [41]
	average	**99.52**	**86.52**
Skin cancer	–	**Datasets**	
		ISIC 2020 accuracy(%)	**MED NODE accuracy(%)**
	AlexNet	95.72 [55],93.10 [56]	82.00 [57],78.11 [58]
	GoogleNet	91.01 [55]	88.00 [58]
	ResNet101	96.15 [55],95.50 [56]	85.00 [57],84.00 [58]
	average	**94.30**	**83.42**

[a] 98.27 [36]: represents the accuracy of 98.27 % from Ref. [36]; other expressions are similar to this.

Among them, TP (True Positive), representing true positive; FN (False Negative), denoting false negative; TN (True Negative), signifying true negative; and FP (False Positive), indicating false positive.

3.3. Results

To showcase the validity of the proposed data cleaning framework MIDC, experiments are conducted on the diagnosis of four kinds of diseases. The experiment is divided into two stages: (1) Data grading experiment: two datasets of the same diseases are divided into high-accuracy D_h and low-accuracy D_l through data grading for data cleaning experiments. The effectiveness of the classification results are demonstrated through internal and external validation experiments; (2) Data cleaning experiment: using MIDC to clean the D_l dataset to obtain the cleaned dataset D_{lc}, and conducting diagnostic experiments on four kinds of diseases based on D_{lc} to obtain the results of data cleaning. The effectiveness of data cleaning is demonstrated through comparative analysis.

3.3.1. Data grading experiment

The experiment is performed on the four kinds of diseases mentioned in section 2.1, and each type of disease corresponds to two datasets (Table 1). For these datasets, graded processing is required. The experiment is constructed by three steps. Step 1: Literature organization; Step 2: Experimental validation; Step 3: Results confirming based on grading rulers.

Step 1: Based on the dataset described in section 2.1, we have chosen the literature on the use of these datasets in recent years, which have high citation rates and journal impact factors, with high credibility. The arranged results are shown in Table 2.

It can be seen from Table 2 that in the two datasets for diagnosis of diabetes retinal disease, the average accuracy of APTOS dataset based on each classic network in the references listed in Table 2 is 95.72 %, and that of Messidor-2 is 77.17 %, the differ of them is 18.55 %. For diagnosing viral pneumonia, the average accuracy of Chest X-Ray Images is 97.99 %, and the average accuracy of COVID-19 is 86.11 %, the differ is 11.88 %. For the diagnosis of breast tumors, the average accuracy of us-data is 99.52 %, and the average accuracy value of BUSI is 86.52 %, the differ is 13.00 %. For the diagnosis of skin cancer, the average accuracy of ISIC 2020 is 94.30 %, and the average accuracy value of MED NODE is 83.42 %, the differ is 10.88 %. Therefore, it can be preliminarily determined that the quality of APTOS, Chest X-Ray Images, us-data, and ISIC 2020 datasets is better, because of their accuracy of dataset label is higher, and therefore they can be set as D_h datasets. The quality of Messidor-2, COVID-19, BUSI, and MED NODE datasets is poor, which means the accuracy of their dataset label is low. Therefore, they can be set as D_l datasets.

We further demonstrate the difference in average accuracy between two datasets of the same kind through Fig. 5. The horizontal axis depicts the type of disease, while the vertical axis showcases the average diagnostic accuracy of the datasets. And based on Fig. 5 the same conclusion as Table 2 can be drawn.

Step 2: Based on D_h and D_l obtained in step 1 and by using the same experimental settings as the literature listed in Table 2, experimental validation process is performed to verify the quality of them. Since the experiments are conducted to compare with existing references, in order to ensure comparability of the results, we use the same network and dataset as those used in existing references for the comparative experiment. The experiments include internal validation and external validation: (1) The internal validation experiment includes the experiment of D_h and D_l. The internal validation experiment of D_h means both training process for diagnose model and test process for diagnose result are performed on D_h, and that of D_l means these processes are both performed on D_l. (2) In the external validation experiment, D_h are set as training set to be trained to obtain the diagnose model, and D_l are set as test data to be put into this diagnose model for the diagnostic result. Table 3 shows the experimental results for the four disease datasets.

From Table 3, it can be seen that for the internal experiments: our experimental results are consistent with the conclusions of step1, which means that the D_h dataset is more accurate, the average accuracies of APTOS, Chest X-Ray Images, us-data, and ISIC 2020 are 95.20 %, 96.56 %, 99.91 %, and 93.33 %, respectively. And the D_l dataset is less accurate, with a lower accuracy of Messidor-2, COVID-19, BUSI, and MED NODE values of 71.18 %, 82.50 %, 85.58 %, and 84.55 %, respectively. For the external experiments, in which the

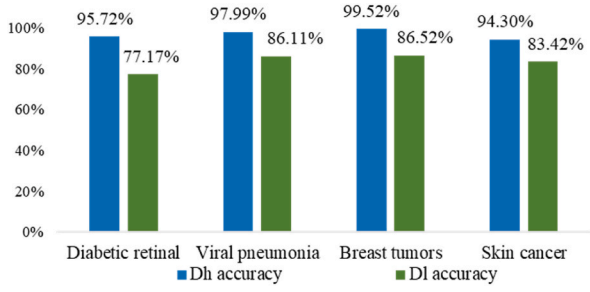


Fig. 5. Difference in average accuracy between the four disease datasets.

Table 3
Experimental results for the four disease datasets.

Disease type	Network	Internal validation		External Validation
		D_h accuracy(%)	D_l accuracy(%)	D_l accuracy(%)
Diabetic retinal (D_h: APTOS	ResNet50	98.22	66.80	73.84
D_l: Messidor-2)	VGG16	91.02	71.25	81.90
	VGG19	93.99	69.63	80.40
	InceptionV3	93.53	70.44	81.90
	InceptionResNetV2	96.04	70.44	79.23
	Xception	97.40	72.07	81.73
	MobileNetV2	96.18	77.63	86.21
	average	**95.20**	**71.18**	**80.74**
Viral pneumonia (D_h: Chest X-Ray Images	VGG16	96.50	82.67	91.33
D_l: COVID-19)	ResNet18	97.09	81.81	92.18
	ResNet50	96.58	83.79	92.64
	DenseNet161	96.07	81.73	92.13
	average	**96.56**	**82.50**	**92.07**
Breast tumors (D_h: us-data	VGG16	99.65	82.36	85.55
D_l: BUSI)	ResNet18	100	86.15	87.22
	ResNet50	100	86.92	88.88
	ResNet101	100	86.92	88.88
	average	**99.91**	**85.58**	**87.63**
Skin cancer (D_h: ISIC 2020	AlexNet	94.07	85.50	88.88
D_l: MED NODE)	GoogleNet	91.11	82.40	84.44
	ResNet101	94.81	85.75	89.62
	average	**93.33**	**84.55**	**87.64**

diagnostic models are trained on D_h and diagnose process are performed on D_l, the average accuracies of above four diseases are 80.74 %, 92.07 %, 87.63 %, and 87.64 %, respectively, which are higher than those obtained from the internal validation of D_l. This shows that, firstly, the network model trained based on D_h is not only better but also has good generalization, indicating that the data quality of D_h is better; secondly, based on the same test set D_l, the model trained with D_h is better than the model trained with D_l, indicating the quality of D_h is better than D_l, i.e. data labels of D_h are more accurate. Except for the literature, the above experiments further demonstrated that the quality of D_h is superior to D_l.

Step3: Based on the grading principle described in section 2.2, we can confirm the grading results: (1) The results of step1 (Table 2) indicate that the accuracy rate of D_h dataset is higher than that of D_l dataset in the reference; (2) The results of step2 (Table 3) indicate that the data quality of Dh dataset is better than that of Dl dataset in the validation experiments. In summary, it can be confirmed that the grading results of D_h and D_l are reasonable and valid.

3.3.2. Data cleaning experiments

Based on the D_h and D_l obtained from above section, data cleaning experiments are performed according to our MIDC framework. In

Table 4
Statistics on changes in the amount of datasets.

Disease type	pre-cleaning(D_l)	post-cleaning(D_{lc})		
		D_{lc1}(method 1)		D_{lc2}(method 2)
Diabetic retinal (D_l:Messidor-2)	1786	ResNet50	1165	1317
		VGG16	1188	
		VGG19	1181	
		InceptionV3	1171	
		InceptionResNetV2	1185	
		Xception	1188	
		MobileNetV2	1199	
Viral pneumonia (D_l:COVID-19)	2950	VGG16	2350	2478
		ResNet18	2338	
		ResNet50	2369	
		DenseNet161	2350	
Breast tumors (D_l:BUSI)	647	VGG16	396	502
		ResNet18	401	
		ResNet50	411	
		ResNet101	411	
Skin cancer (D_l: MED NODE)	680	AlexNet	457	483
		GoogleNet	465	
		ResNet101	470	

* method 1:multi-training; method 2:multi-network.

the experiments, two diagnose methods, multi-training and multi-network methods are applied respectively. After data cleaning, the error data of D_l are deleted and the change of its data amount are shown in Table 4.

In Table 4, D_{lc1} is the cleaned dataset obtained by using method 1 and D_{lc2} is obtained by using method 2. From the table, it can be seen that (1) D_{lc1} and D_{lc2} are both lower than D_l, which means the size of D_l has decreased during data cleaning procedure; (2) D_{lc1} is lower than D_{lc2}, which means that the amount of data cleaned by method2 is lower than that of method1. Taking Messidor-2 as an example: the amount of data cleaned by multi-training based on MobileNetV2 network changed from 1786 to 1199, i.e., 587 error data are cleaned, while 469 error data are cleaned by the multi-network method. Fig. 6 shows the proportion of error data being deleted and correct data being remained in the datasets of four diseases.

Combining Figs. 5 and 6, it is evident that for the Messidor-2 dataset, based on which the accuracy is 18.55 % lower than that of APTOS, has 469 error data, which is about 26.26 % of the total; for COVID-19 dataset, the accuracy is 11.88 % lower than that of Chest X-Ray Images, the error data is 472,which is about 16.00 % of the total; for BUSI dataset, the accuracy is 13.00 % lower than that of us-data, the error data is 145, which is about 23.50 % of the total number, and for MED NODE dataset, the accuracy is 10.88 % lower than that of ISIC 2020, the error data is 197, which is about 28.97 % of the total number. According to the above statement we can see that the percentage of error data in D_l is in proportion to its differ value compared to D_h in terms of accuracy.

By putting the cleaned data D_{lc1} and D_{lc2} into the networks presented in Table 2 respectively, the diagnose results of cleaned dataset are obtained. Table 5 presents the different diagnostic results of D_{lc1} and D_{lc2}.

As can be seen from Tables 5 and in all networks, the results based on D_{lc2} are higher than those based on D_{lc1} dataset, which means that cleaning with multi-network is better than cleaning with multi-training. In Messidor-2, the accuracy of ResNet50 network based on D_{lc2} is 81.53 %, while the accuracy based on D_{lc1} dataset is 79.23 %. In COVID-19, the accuracy of VGG16 network based on D_{lc2} dataset is 93.68 %, while that of D_{lc1} dataset is 91.36 %. In BUSI, the accuracy of the VGG16 network based on the D_{lc2} dataset is 89.86 %, while that of D_{lc1} dataset is 86.92 %. In MED NODE, the accuracy of the AlexNet network based on the D_{lc2} dataset is 94.07 %, while that of D_{lc1} dataset is 91.11 %.

To present the efficiency of our data cleaning framework, Table 6 compares the diagnostic results before and after data cleaning, as well as the diagnostic results of these datasets in the references. Where the test set used for all experiments is D_l, but their training sets used for the training of diagnostic models are different.

From Table 6, we can see: (1) Based on the datasets without data cleaning, i.e.,D_l and D_h, the accuracy rate of the model trained on D_h is higher than that of D_l; (2) Based on the datasets being cleaned by MIDC framework, i.e.,D_{lc1} and D_{lc2}, both the accuracy rates of the model trained on them are higher than that of D_l and D_h; (3) The accuracy rate of the model trained on D_{lc2} is higher than that of D_{lc1}. (4) The accuracy after cleaning is higher than the average accuracy of the reference literature. Take diabetic retinal as an example, the average reference accuracy on D_l is 77.17 %, the average diagnostic accuracy of the model trained on D_l is 71.18 %, the average diagnostic accuracy of the model trained on D_h is 80.74 %, the average diagnostic accuracy of the model trained on D_{lc1} is 83.26 %, while that of D_{lc2} reaches 85.13 %.

In order to more directly show the change of accuracy before and after cleaning, we show the experimental results of the four types of datasets through Fig. 7, and it can be seen that for four kinds of datasets there is a relatively large increase in the accuracy after cleaning.

To further illustrate the performance of our method, the confusion matrix results before and after cleaning shown in Fig. 8.

As can be seen in Fig. 8, columns (a)–(d) represents different training datasets, which are (a) D_l training, (b) D_h training, (c) $D_{lc\text{-}1}$ training, and (d) $D_{lc\text{-}2}$ training. Lines A-D represents datasets of different disease types, which are A. Diabetic retinal dataset (Messidor-2), B. Viral pneumonia dataset (COVID-19), C. Breast tumors dataset (BUSI), and D. Skin cancer (MED NODE). For all four datasets the number of correctly classified images is better than that of before cleaning. Take diabetic retinal Messidor-2 dataset as an example, the number of correctly classified images of the model trained on Dl is 235, the number of correctly classified images of the model trained on D_h is 289, the number of correctly classified images of the model trained on D_{lc1} is 298, while that of D_{lc2} reaches 305.

3.3.3. *Comparison experiment with existing method*

This experiment compares the proposed framework and existing method to further validate the performance of the MIDC framework. Currently, only Li et al. [21] perform data cleaning based on medical images, and since the early esophageal cancer dataset they use is private, we can only make a rough comparison by applying their method to the public dataset. We compare the amount and

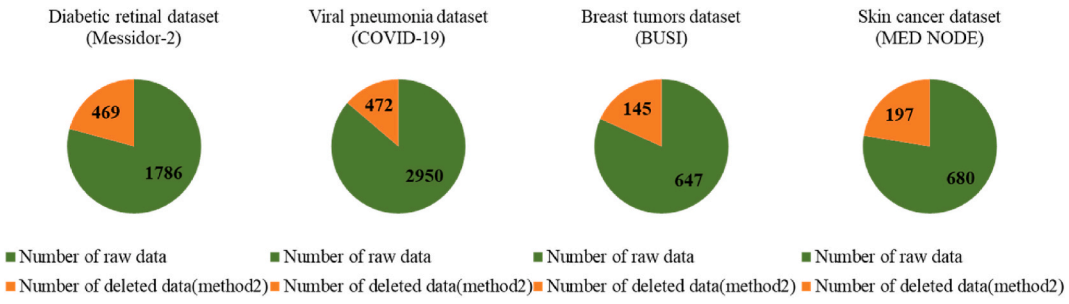


Fig. 6. Percentage of the error data to the total.

Table 5
Experimental results based on D_{lc1} and D_{lc2} with different networks.

Dataset	Network	D_{lc1}				D_{lc2}			
		Accuracy (%)	Recall (%)	Precision (%)	F1 (%)	Accuracy (%)	Recall (%)	Precision (%)	F1 (%)
Messidor-2	ResNet50	79.23	79.23	78.92	78.95	**81.53**	81.53	82.21	81.79
	VGG16	83.68	83.68	86.90	85.08	**85.25**	85.25	85.71	85.34
	VGG19	82.26	82.26	81.94	82.10	**84.33**	84.33	84.74	84.42
	InceptionV3	80.85	80.85	80.49	80.66	**83.55**	83.55	83.22	83.34
	InceptionResNetV2	83.07	83.07	82.93	82.81	**85.10**	85.10	85.10	85.10
	Xception	83.68	83.68	85.78	84.62	**84.39**	84.39	84.39	84.39
	MobileNetV2	90.07	90.07	89.60	89.72	**91.73**	91.73	91.73	91.73
COVID-19	VGG16	91.36	91.36	91.67	91.40	**93.68**	93.68	93.69	93.68
	ResNet18	90.94	90.94	91.78	91.02	**93.89**	93.89	93.98	93.90
	ResNet50	91.61	91.61	91.65	91.61	**93.68**	93.68	93.75	93.69
	DenseNet161	90.31	90.31	90.82	90.37	**93.89**	93.89	93.97	93.88
BUSI	VGG16	86.92	86.92	87.01	86.96	**89.86**	89.86	89.85	89.85
	ResNet18	91.42	91.42	91.50	91.44	**93.05**	93.05	93.06	93.05
	ResNet50	92.00	92.00	92.27	92.02	**95.45**	95.45	99.09	96.91
	ResNet101	92.00	92.00	92.27	92.27	**95.45**	95.45	99.09	96.91
MED NODE	AlexNet	91.11	91.11	91.66	91.05	**94.07**	94.07	94.09	94.06
	GoogleNet	90.37	90.37	90.53	90.37	**93.75**	93.75	93.88	93.75
	ResNet101	91.85	91.85	92.03	91.81	**94.81**	94.81	94.87	94.06

Table 6
Experimental results before and after cleaning.

Disease type	Network	Average references accuracy(%)	Pre-cleaning dataset training accuracy(%)		Post-cleaning dataset training accuracy(%)	
		D_l	D_l	D_h	D_{lc1}	D_{lc2}
Diabetic retinal (D_h: APTOS D_l: Messidor-2)	ResNet50	74.52	66.80	73.84	79.23	81.53
	VGG16	80.46	71.25	81.90	83.68	85.25
	VGG19	79.46	69.63	80.40	82.26	84.33
	InceptionV3	73.44	70.44	81.90	80.85	83.55
	InceptionResNetV2	72.87	70.44	79.23	83.07	85.10
	Xception	78.77	72.07	81.73	83.68	84.39
	MobileNetV2	83.81	77.63	86.21	90.07	91.73
	average	**77.17**	**71.18**	**80.74**	**83.26**	**85.13**
Viral pneumonia (D_h: Chest X-Ray Images D_l: COVID-19)	VGG16	84.49	82.67	91.33	91.36	93.68
	ResNet18	85.78	81.81	92.18	90.94	93.89
	ResNet50	92.63	83.79	92.64	91.61	93.68
	DenseNet161	83.52	81.73	92.13	90.31	93.89
	average	**86.11**	**82.50**	**92.07**	**91.06**	**93.79**
Breast tumors (D_h: us-data D_l: BUSI)	VGG16	85.76	82.36	85.55	86.92	89.86
	ResNet18	86.65	86.15	87.22	91.42	93.05
	ResNet50	87.45	86.92	88.88	92.00	95.45
	ResNet101	86.05	86.92	88.88	92.00	95.45
	average	**86.52**	**85.59**	**87.63**	**90.59**	**93.45**
Skin cancer (D_h: ISIC 2020 D_l: MED NODE)	AlexNet	80.06	85.50	88.88	91.11	94.07
	GoogleNet	88.00	82.40	84.44	90.37	93.75
	ResNet101	84.50	85.75	89.62	91.85	94.81
	average	**83.42**	**84.55**	**87.64**	**91.11**	**94.21**

the accuracy of datasets obtained after cleaning in two methods. To ensure comparable results, we used the same networks and datasets as described above section, the experimental results are shown in Table 7.

From Table 7, we can see: (1) For the amount of cleaned datasets, the amount of datasets cleaned by Ref. [21] are all less than the amount of datasets cleaned by the MIDC framework. (2) For the accuracy of the cleaned datasets, the accuracy of the datasets cleaned by the MIDC framework are all higher than the accuracy of the datasets cleaned by Ref. [21]. Take diabetic retinal as an example, the amount of the datasets cleaned by the MIDC framework is 1317, the amount of the datasets cleaned by Ref. [21] is 935. the average accuracy of the datasets cleaned by the MIDC framework is 85.13 %, the average accuracy of the datasets cleaned by Ref. [21] is 77.57 %.

4. Discussion

To illustrate the effectiveness of the proposed medical image dataset cleaning framework MIDC, we conduct experiments on four diseases, which are diabetic retina, viral pneumonia, breast tumors, and skin cancer. The experiments includes three stages: data

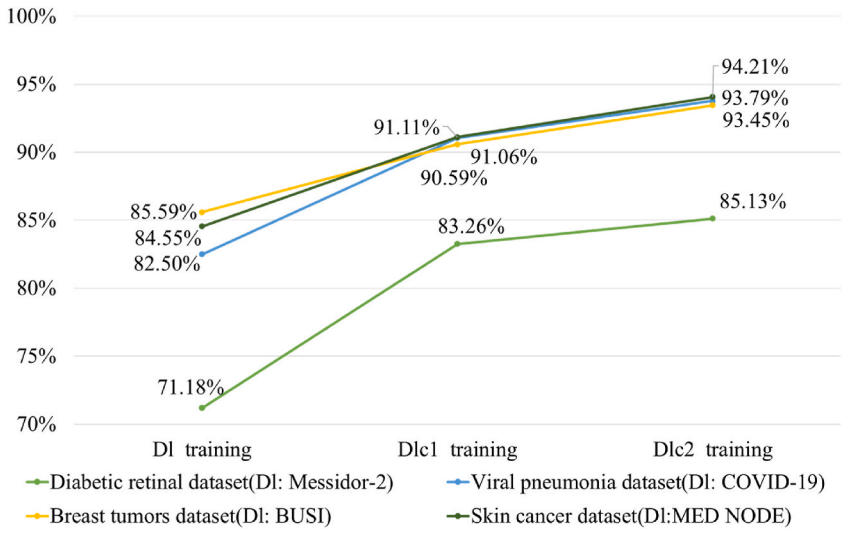


Fig. 7. Change in accuracy before and after cleaning.

grading experiments, data cleaning experiments, and comparison experiment with existing method.

In the data grading experiments, the two datasets of each disease are divided into high-accuracy dataset D_h and low-accuracy dataset D_l according to the designed rules of data grading. Table 2 lists the results of references in recent years for the datasets of these disease, and according to the different accuracy of these datasets, they are divided into D_h and D_l. To further verify the quality of D_h and D_l, validation experiments are performed. The experimental results are shown in Table 3, which demonstrated that the diagnostic result of D_h is more accurate than that of D_l.

Based on D_h and D_l, the data cleaning experiments are performed to testify the efficiency of the proposed MIDC. According the framework, which clean the low-accuracy dataset D_l by high-accuracy dataset D_h, a cleaned dataset D_{lc} are obtained. As two diagnostic methods, multi-training and multi-network, are designed in the framework, two kinds of cleaned datasets are obtained, D_{lc1} and D_{lc2}. Based on the experimental results, following conclusion can be drawn: (1) After data cleaning, the change of D_l data amount is shown in Table 4, which shows the error data in D_l is deleted and the number of deleted data by multi-training method is higher than that of multi-network method. (2) Table 5 presents the experimental results based on D_{lc1} and D_{lc2} with different networks. In the table we can see that the accuracy of diagnostic results on D_{lc2} is higher than that of D_{lc1}, which means the quality of D_{lc2} is better than D_{lc1}, i.e., multi-network method performs better than multi-training method. The reason for different performance is that in the multi-training methods, the same network has limitations in extracting features, which leads to the same features being extracted even in multiple training. While for multi-network method, different networks can extract different features that make up the defects of multi-training method. (3) In Table 6, the diagnostic results before and after data cleaning are compared. In the table, the diagnostic results based on the datasets being cleaned is better than that of dataset without data cleaning, which testified the efficiency of our MIDC framework.

In the comparison experiment, by comparing with the existing method, the performance of the proposed MIDC framework is further validated. Table 7 lists the results of accuracy and amount of datasets obtained after cleaning in two methods. From Tables 7 and it can be seen that the amount of data cleaned by the comparison method is less than the amount of data cleaned by the MIDC framework, and over-cleaning of the dataset also leads to a decrease in accuracy. The proposed MIDC framework directly selects the data based on the labels of the data rather than relying only on the distribution of the data features, which reduces the risk of the correct data being misclassified as outliers. From Tables 7 and it can be seen that the results of the proposed framework are all better than the results of the comparison method.

5. Conclusion

In this paper, we propose a medical image dataset cleaning framework MIDC based on deep learning. This framework uses high-accuracy datasets to clean low-accuracy datasets. In the experiments, we use four couples of datasets that belonging four kinds of diseases respectively to demonstrate the effectiveness of the proposed framework. Among them, for the APTOS dataset of diabetic retinal, the average accuracy after cleaning increased from 71.18 % to 85.13 %. For the Chest X-Ray Images dataset of virtual pneumonia, the average accuracy after cleaning increased from 82.50 % to 93.79 %. For the BUSI dataset of breast tumors, the average accuracy after cleaning increased from 85.59 % to 93.45 %. For the MED NODE dataset of skin cancer, the average accuracy after cleaning increased from 84.55 % to 94.21 %. The experimental results indicate that the classification accuracy is significantly improved after data cleaning compared to before cleaning.

The proposed MIDC framework in this paper automates the cleaning of medical image data to improve the quality of public datasets. This process does not rely on annotations from professional doctors, nor does it require additional datasets with more reliable labels. By introducing the MIDC framework, the beneficiaries include: (1) Professional physicians, for reliable computer-aided

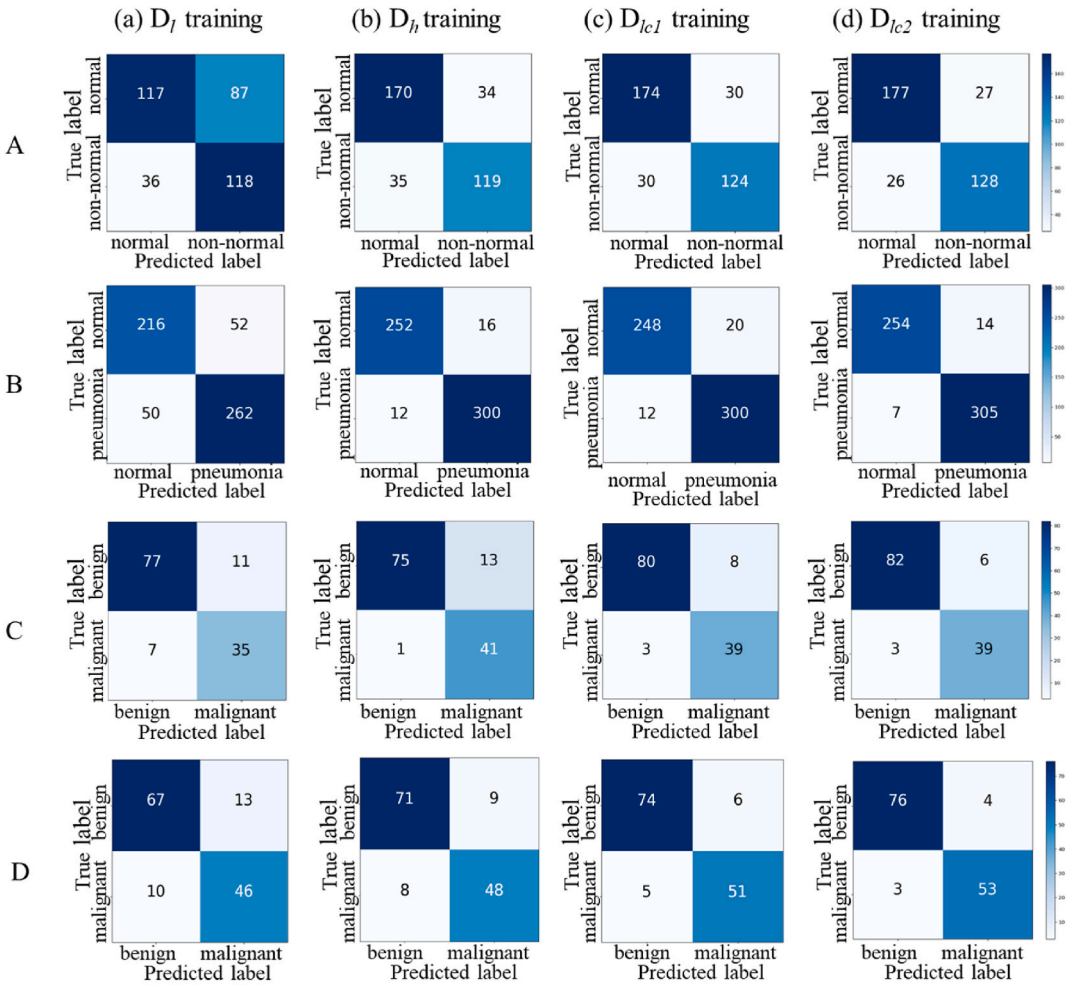


Fig. 8. Confusion matrices before and after cleaning. Columns (a)–(d): (a) D_l training, (b) D_h training, (c) D_{lc-1} training, (d) D_{lc-2} training. Lines A–D: A. Diabetic retinal dataset (Messidor-2), B. Viral pneumonia dataset (COVID-19), C. Breast tumors dataset (BUSI), D. Skin cancer (MED NODE).

diagnosis results can better assist them in making diagnoses and reduce misdiagnosis rates. (2) Biomedical engineers, higher quality data enables more reliable research results and innovation in diagnostic techniques. (3) Patients, more accurate and early diagnosis leads to better prognosis and more timely intervention.

However, there are still some limitations to this work, as it currently only cleans datasets for four diseases and cannot cover all clinical application scenarios. The existing cleaning algorithms still need to be optimized to improve higher accuracy. In future work, we will gradually address the limitations of this work. Firstly, we will optimize the data grading module and data cleaning module, and introduce more advanced deep learning algorithms; Secondly, we will integrate more datasets and disease types to further validate the generalization ability of the framework; Then, we plan to explore the integration of advanced image compression techniques into the framework to further optimize the data processing flow; Finally, improve the real-time application and applicability of the framework in practical clinical environments.

Ethics declarations

Review and/or approval by an ethics committee was not needed for this study because the data used in this study were all publicly available datasets.

Data availability statement

The datasets supporting this study are derived from available to the public. (1) Diabetic retinal: APTOS-https://doi.org/10.5455/

Table 7
Comparison experimental results with existing method.

Dataset	Network	$D_{lc\text{-}}$ [21]					$D_{lc\text{-}MIDC}$				
		Accuracy (%)	Recall (%)	Precision (%)	F1 (%)	amount	Accuracy (%)	Recall (%)	Precision (%)	F1 (%)	amount
Messidor-2 (original amount: 1786)	ResNet50	72.72	72.72	72.28	72.33	935	81.53	81.53	82.21	81.79	1317
	VGG16	79.14	79.14	79.23	79.18		85.25	85.25	85.71	85.34	
	VGG19	78.07	78.07	78.17	78.12		84.33	84.33	84.74	84.42	
	InceptionV3	77.00	77.00	77.32	76.48		83.55	83.55	83.22	83.34	
	InceptionResNetV2	74.86	74.86	74.50	74.50		85.10	85.10	85.10	85.10	
	Xception	77.54	77.54	77.30	77.18		84.39	84.39	84.39	84.39	
	MobileNetV2	83.68	83.68	86.90	85.08		91.73	91.73	91.73	91.73	
	average	77.57	77.57	77.96	77.55		**85.13**	**85.13**	**85.30**	**85.16**	
COVID-19 (original amount: 2950)	VGG16	92.58	92.58	92.86	92.60	2090	93.68	93.68	93.69	93.68	2478
	ResNet18	90.67	90.67	90.89	90.67		93.89	93.89	93.98	93.90	
	ResNet50	91.38	91.38	91.47	91.40		93.68	93.68	93.75	93.69	
	DenseNet161	91.38	91.38	91.47	91.40		93.89	93.89	93.97	93.88	
	average	91.50	91.50	91.67	91.52		**93.79**	**93.79**	**93.85**	**93.79**	
BUSI (original amount: 647)	VGG16	83.72	83.72	79.60	79.72	389	89.86	89.86	89.85	89.85	502
	ResNet18	90.94	90.94	91.78	91.02		93.05	93.05	93.06	93.05	
	ResNet50	93.02	93.02	93.53	91.66		95.45	95.45	99.09	96.91	
	ResNet101	93.02	93.02	93.53	91.66		95.45	95.45	99.09	96.91	
	average	90.18	90.18	89.61	88.52		**93.45**	**93.45**	**95.28**	**94.18**	
MED NODE (original amount: 680)	AlexNet	87.50	87.50	88.80	88.15	358	94.07	94.07	94.09	94.06	483
	GoogleNet	81.90	81.90	82.25	81.95		93.75	93.75	93.88	93.75	
	ResNet101	87.50	87.50	87.54	87.50		94.81	94.81	94.87	94.06	
	average	85.63	85.63	86.20	85.87		**94.21**	**94.21**	**94.28**	**93.96**	

* $\mathbf{D_{lc\text{-}}}$ [21] represents the D_l dataset cleaned by the method of [21]; $\mathbf{D_{lc\text{-}MIDC}}$ represents the D_l dataset cleaned by the method of MIDC.

aim.2019.27.327-332, Messidor-2-https://doi.org/10.5566/ias.1155. (2) Viral pneumonia: Chest X-Ray Images-https://doi.org/10.17632/m4s2jn3csb.1, COVID-19-https://doi.org/10.1109/ACCESS.2020.3010287. (3) Breast tumors: us-data-https://doi.org/10.17632/wmy84gzngw.1, BUSI-https://doi.org/10.1016/j.dib.2019.104863. (4) Skin cancer: ISIC 2020-https://doi.org/10.1038/s41597-021-00815-z, MED NODE-https://doi.org/10.1016/j.eswa.2015.04.034. We have provided data links in the manuscript.

CRediT authorship contribution statement

Sanli Yi: Writing – review & editing, Writing – original draft, Validation, Resources, Project administration, Methodology, Funding acquisition, Data curation, Conceptualization. **Ziyan Chen:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by The National Natural Science Foundation of China, Regional Science Foundation Project(No. 62266025).

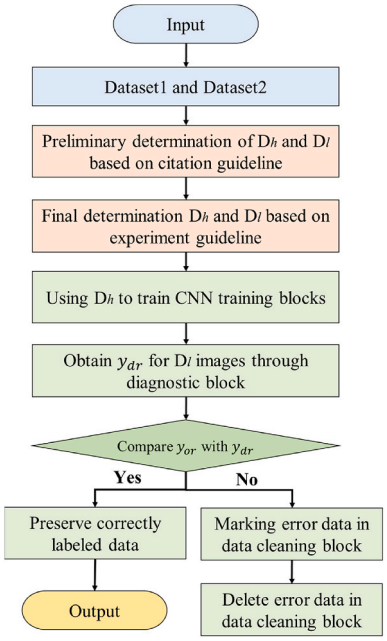
Appendix A

Algorithm 1. MIDC Framework for Dataset Cleaning

Input: Two datasets of the same disease - Dataset1 and Dataset2
Output: Cleaned dataset D_{lc}

1: Load Datasets
    obtain Dataset1 and Dataset2 through specific guidelines
2: Data Grading Module
    D_h, D_l = data_grading_rules(Dataset1, Dataset2)
3: Data Cleaning Module
    diagnose_model = train_cnn(D_h)
    diagnostic_results = compare(diagnose_model, D_l)
    D_{lc} = clean_dataset(D_l, diagnostic_results)
4: Output Cleaned Dataset
    Output D_{lc}

Appendix B



Appendix C

Algorithm 2. Data cleaning module

Input: Two datasets of the same disease - D_h and D_l
Output: Cleaned dataset D_{lc}

obtain D_h and D_l through data grading module
training CNN with D_h: diagnose_model = train_cnn(D_h)
utilize diagnose_model to diagnose D_l through Eq. (2)
calculate the weighted averaging result using Eq. (1) or Eq. (3)
the final diagnostic result are obtained by Eq. (4)
for i in range(len(original_labels)):
if original_labels[i] ! = final_diagnostic_labels[i]:
delete images[i]
continue
end if
end for 12: D_{lc} = cleaned_images.append(images[i])
Return D_{lc}

References

[1] M.I. Razzak, S. Naz, A. Zaib, Deep learning for medical image processing: overview, challenges and the future, Classification in BioApps: Automation of decision making (2018) 323–350, https://doi.org/10.1007/978-3-319-65981-7_12.
[2] H. Li, Deep learning for natural language processing: advantages and challenges, Natl. Sci. Rev. 5 (1) (2018) 24–26, https://doi.org/10.1093/NSR/NWX110.
[3] H. Al-Khazraji, A.R. Nasser, A.M. Hasan, A.K. Al Mhdawi, H. Al-Raweshidy, A.J. Humaidi, Aircraft engines remaining useful life prediction based on a hybrid model of autoencoder and deep belief network, IEEE Access 10 (2022) 82156–82163, https://doi.org/10.1109/ACCESS.2022.3188681.
[4] R.H. Hadi, H.N. Hady, A.M. Hasan, A. Al-Jodah, A.J. Humaidi, Improved fault classification for predictive maintenance in industrial IoT based on AutoML: a case study of ball-bearing faults, Processes 11 (5) (2023) 1507, https://doi.org/10.3390/pr11051507.
[5] A.R. Nasser, A.M. Hasan, A.J. Humaidi, DL-AMDet: deep learning-based malware detector for android, Intelligent Systems with Applications 21 (2024) 200318, https://doi.org/10.1016/j.iswa.2023.200318.
[6] D. Shen, G. Wu, H.I. Suk, Deep learning in medical image analysis, Annu. Rev. Biomed. Eng. 19 (2017) 221–248, https://doi.org/10.1146/annurev-bioeng-071516-044442.

[7] Y. Gong, G. Liu, Y. Xue, R. Li, L. Meng, A survey on dataset quality in machine learning, Inf. Software Technol. 162 (2023), https://doi.org/10.1016/j.infsof.2023.107268.

[8] C. Sager, C. Janiesch, P. Zschech, A survey of image labelling for computer vision applications, in: Journal of Business Analytics, 2021, https://doi.org/10.1080/2573234X.2021.1908861.

[9] C. Qian, B. Huang, X. Yang, G. Chen, Data science for oceanography: from small data to big data, Big Earth Data 6 (2) (2022) 236–250, https://doi.org/10.1080/20964471.2021.1902080.

[10] T. Xu, Y. Xu, S. Yang, B. Li, W. Zhang, Learning accurate label-specific features from partially multilabeled data, IEEE Transact. Neural Networks Learn. Syst. (2023), https://doi.org/10.1109/TNNLS.2023.3241921.

[11] D. Karimi, H. Dou, S.K. Warfield, A. Gholipour, Deep learning with noisy labels: exploring techniques and remedies in medical image analysis, Med. Image Anal. 65 (2020), https://doi.org/10.1016/j.media.2020.101759.

[12] H. Galhardas, D. Florescu, D. Shasha, E. Simon, AJAX: an extensible data cleaning tool, in: ACM SIGMOD Conference, 2000, https://doi.org/10.1145/342009.336568.

[13] J. Brownlee, Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python, Machine Learning Mastery, 2020.

[14] K. Mavrogiorgos, A. Kiourtis, A. Mavrogiorgou, S. Kleftakis, D. Kyriazis, A multi-layer approach for data cleaning in the healthcare domain, in: Proceedings of the 2022 8th International Conference on Computing and Data Engineering, 2022, January, pp. 22–28. https://dl.acm.org/doi/abs/10.1145/3512850.3512856.

[15] S. Dong, P. Wang, K. Abbas, A survey on deep learning and its applications, Comput. Sci. Rev. 40 (2021) 100379, https://doi.org/10.1016/j.cosrev.2021.100379.

[16] Y. Roh, G. Heo, S.E. Whang, A survey on data collection for machine learning: a big data - ai integration perspective, IEEE Trans. Knowl. Data Eng. 33 (4) (2021) 1328–1347, https://doi.org/10.1109/tkde.2019.2946162.

[17] Y. Zhang, Z. Jin, F. Liu, W. Zhu, W. Mu, W. Wang, ImageDC: image data cleaning framework based on deep learning, in: 2020 IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS), 2020, pp. 748–752, https://doi.org/10.1038/s41598-022-18707-6.

[18] D. Liu, Y. Meng, L. Wang, Data cleaning of irrelevant images based on transfer learning, in: Proceedings - 2020 International Conference on Intelligent Computing, Automation and Systems, 2020, pp. 450–456, https://doi.org/10.1109/ICICAS51530.2020.00099. ICICAS 2020.

[19] T. Kavzoglu, Increasing the accuracy of neural network classification using refined training data, Environ. Model. Software 24 (7) (2009) 850–858, https://doi.org/10.1016/j.envsoft.2008.11.012.

[20] C.G. Northcutt, L. Jiang, I.L. Chuang, Confident learning: estimating uncertainty in dataset labels, J. Artif. Intell. Res. 70 (2019) 1373–1411, https://doi.org/10.1613/jair.1.12125.

[21] Z. Li, R. Wu, T. Gan, Study on image data cleaning method of early esophageal cancer based on VGG_NIN neural network, Sci. Rep. 12 (1) (2022), https://doi.org/10.1038/s41598-022-18707-6.

[22] Sruthy Manmadhan, Binsu C. Kovoor, Visual question answering: a state-of-the-art review, Artif. Intell. Rev. 53 (8) (2020) 5705–5745, https://doi.org/10.1007/s10462-020-09832-7, 2020.

[23] K. Adane, M. Gizachew, S. Kendie, The role of medical data in efficient patient care delivery: a review, in: Risk Management and Healthcare Policy, vol. 12, Dove Medical Press Ltd, 2019, pp. 67–73, https://doi.org/10.2147/RMHP.S179259.

[24] M.D.P. Raj Vincent, M. Aamir, J. Röglin, Improving classification results on a small medical dataset using a GAN; an outlook for dealing with rare disease datasets, Front. Comput. Sci. (2022), https://doi.org/10.3389/fcomp.2022.858874 (n.d.).

[25] A. Guo, P. Wang, The current state of doctors' communication skills in mainland China from the perspective of doctors' self-evaluation and patients' evaluation: a cross-sectional study, Patient Educ. Counsel. 104 (7) (2021) 1674–1680, https://doi.org/10.1016/j.pec.2020.12.013.

[26] L. Oakden-Rayner, Exploring large-scale public medical image datasets, Acad. Radiol. 27 (1) (2020) 106–112, https://doi.org/10.1016/j.acra.2019.10.006.

[27] N.E.M. Khalifa, M. Loey, M.H.N. Taha, H.N.E.T. Mohamed, Deep transfer learning models for medical diabetic retinopathy detection, Acta Inf. Med. : AIM : journal of the Society for Medical Informatics of Bosnia & Herzegovina : casopis Drustva za medicinsku informatiku BiH 27 (5) (2019) 327–332, https://doi.org/10.5455/aim.2019.27.327-332.

[28] E. Decencière, X. Zhang, G. Cazuguel, B. Laÿ, B. Cochener, C. Trone, J.C. Klein, Feedback on a publicly distributed image database: the Messidor database, Image Anal. Stereol. 33 (3) (2014) 231–234, https://doi.org/10.5566/ias.1155.

[29] J. Cuadros, G. Bresnick, EyePACS: an adaptable telemedicine system for diabetic retinopathy screening, J. Diabetes Sci. Technol. 3 (3) (2009) 509–516, https://doi.org/10.1177/193229680900300315.

[30] K.M. Almustafa, A.K. Sharma, S. Bhardwaj, STARC: deep learning Algorithms' modelling for STructured analysis of retina classification, Biomed. Signal Process Control 80 (2023) 104357, https://doi.org/10.1016/j.bspc.2022.104357.

[31] Md Alamin Talukder, Chest X-Ray Image, Mendeley Data, V1, 2023, https://doi.org/10.17632/m4s2jn3csb.1.

[32] M.E.H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M.A. Kadir, Z.B. Mahbub, M.T. Islam, Can AI help in screening viral and COVID-19 pneumonia? IEEE Access 8 (2020) 132665–132676, https://doi.org/10.1109/ACCESS.2020.3010287.

[33] Paulo Sergio Rodrigues, Breast Ultrasound Image, Mendeley Data, V1, 2018, https://doi.org/10.17632/wmy84gzngw.1. https://data.mendeley.com/datasets/wmy84gzngw/1.

[34] W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy, Dataset of breast ultrasound images, Data Brief 28 (2019) 104863, https://doi.org/10.1016/j.dib.2019.104863.

[35] Y. Benhammou, B. Achchab, F. Herrera, S. Tabik, BreakHis based breast cancer automatic diagnosis using deep learning: taxonomy, survey and insights, Neurocomputing 375 (2020) 9–24, https://doi.org/10.1016/j.neucom.2019.09.044.

[36] G. Yue, Y. Li, T. Zhou, X. Zhou, Y. Liu, T. Wang, Attention-driven cascaded network for diabetic retinopathy grading from fundus images, Biomed. Signal Process Control 80 (2023), https://doi.org/10.1016/j.bspc.2022.104370.

[37] C. Lahmar, A. Idri, On the value of deep learning for diagnosing diabetic retinopathy, Health Technol. 12 (1) (2022) 89–105, https://doi.org/10.1007/s12553-021-00606-x.

[38] H. Li, N. Zeng, P. Wu, K. Clawson, Cov-Net: a computer-aided diagnosis method for recognizing COVID-19 from chest X-ray images via machine vision, Expert Syst. Appl. 207 (2022), https://doi.org/10.1016/j.eswa.2022.118029.

[39] A.M. Ismael, A. Şengür, Deep learning approaches for COVID-19 detection based on chest X-ray images, Expert Syst. Appl. 164 (2021), https://doi.org/10.1016/j.eswa.2020.114054.

[40] M. Masud, M.S. Hossain, H. Alhumyani, S.S. Alshamrani, O. Cheikhrouhou, S. Ibrahim, G. Muhammad, A.E.E. Rashed, B.B. Gupta, Pre-trained convolutional neural networks for breast cancer detection using ultrasound images, ACM Trans. Internet Technol. 21 (4) (2021), https://doi.org/10.1145/3418355.

[41] W.K. Moon, Y.W. Lee, H.H. Ke, S.H. Lee, C.S. Huang, R.F. Chang, Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks, Comput. Methods Progr. Biomed. 190 (2020) 105361, https://doi.org/10.1016/j.cmpb.2020.105361.

[42] A. Paullada, I.D. Raji, E.M. Bender, E. Denton, A. Hanna, Data and its (dis)contents: a survey of dataset development and use in machine learning research, Patterns (New York, N.Y.) 2 (11) (2021) 100336, https://doi.org/10.1016/j.patter.2021.100336.

[43] H.Y. Teh, A.W. Kempa-Liehr, K.I.K. Wang, Sensor data quality: a systematic review, J Big Data 7 (2020) 11, https://doi.org/10.1186/s40537-020-0285-1.

[44] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman, A. Halpern, B. Helba, H. Kittler, K. Kose, S. Langer, K. Lioprys, J. Malvehy, S. Musthaq, J. Nanda, O. Reiter, H.P. Soyer, A patient-centric dataset of images and metadata for identifying melanomas using clinical context, Sci. Data 8 (1) (2021) 34, https://doi.org/10.1038/s41597-021-00815-z.

[45] I. Giotis, N. Molders, S. Land, M. Biehl, M.F. Jonkman, N. Petkov, MED-NODE: a computer-assisted melanoma diagnosis system using non-dermoscopic images, Expert Syst. Appl. 42 (19) (2015) 6578–6585, https://doi.org/10.1016/j.eswa.2015.04.034.

[46] T. Liu, H. Zhang, H. Long, J. Shi, Y. Yao, Convolution neural network with batch normalization and inception-residual modules for Android malware classification, Sci. Rep. 12 (1) (2022) 13996, https://doi.org/10.1038/s41598-022-18402-6.

[47] A.A. Afuwape, Y. Xu, J.H. Anajemba, G. Srivastava, Performance evaluation of secured network traffic classification using a machine learning approach, Comput. Stand. Interfac. 78 (2021) 103545, https://doi.org/10.1016/j.csi.2021.103545.
[48] Z. Lu, J. Miao, J. Dong, S. Zhu, X. Wang, J. Feng, Automatic classification of retinal diseases with transfer learning-based lightweight convolutional neural network, Biomed. Signal Process Control 81 (2023), https://doi.org/10.1016/j.bspc.2022.104365.
[49] S.L. Yi, X.L. Yang, T.W. Wang, F.R. She, X. Xiong, J.F. He, Diabetic retinopathy diagnosis based on RA-efficientnet, Appl. Sci. 11 (22) (2021), https://doi.org/10.3390/app112211035.
[50] H. Mustafa, S.F. Ali, M. Bilal, M.S. Hanif, Multi-stream deep neural network for diabetic retinopathy severity classification under a boosting framework, IEEE Access 10 (2022) 113172–113183, https://doi.org/10.1109/ACCESS.2022.3217216.
[51] G. Jia, H.K. Lam, Y. Xu, Classification of COVID-19 chest X-Ray and CT images using a type of dynamic CNN modification method, Comput. Biol. Med. 134 (2021), https://doi.org/10.1016/j.compbiomed.2021.104425.
[52] A. Paul, A. Basu, M. Mahmud, M.S. Kaiser, R. Sarkar, Inverted bell-curve-based ensemble of deep learning models for detection of COVID-19 from chest X-rays, Neural Comput. Appl. 35 (22) (2023) 16113–16127, https://doi.org/10.1007/s00521-021-06737-6.
[53] A. Raza, N. Ullah, J.A. Khan, M. Assam, A. Guzzo, H. Aljuaid, DeepBreastCancerNet: a novel deep learning model for breast cancer detection using ultrasound images, Appl. Sci. 13 (4) (2023), https://doi.org/10.3390/app13042082.
[54] J. Wang, Y. Zheng, J. Ma, X. Li, C. Wang, J. Gee, H. Wang, W. Huang, Information bottleneck-based interpretable multitask network for breast cancer classification and segmentation, Med. Image Anal. 83 (2023), https://doi.org/10.1016/j.media.2022.102687.
[55] F. Alenezi, A. Armghan, K. Polat, A multi-stage melanoma recognition framework with deep residual neural network and hyperparameter optimization-based decision support in dermoscopy images, Expert Syst. Appl. 215 (2023) 119352, https://doi.org/10.1016/j.eswa.2022.119352.
[56] A. Naeem, T. Anees, M. Fiza, R.A. Naqvi, S.W. Lee, SCDNet: a deep learning-based framework for the multiclassification of skin cancer using dermoscopy images, Sensors 22 (15) (2022) 5652, https://doi.org/10.3390/s22155652.
[57] L. Zhang, H.J. Gao, J. Zhang, B. Badami, Optimization of the convolutional neural networks for automatic detection of skin cancer, Open medicine (Warsaw, Poland) 15 (2020) 27–37, https://doi.org/10.1515/med-2020-0006.
[58] K.M. Hosny, M.A. Kassem, M.M. Foaud, Skin melanoma classification using ROI and data augmentation with deep convolutional neural networks, Multimed. Tool. Appl. 79 (2020) 24029–24055, https://doi.org/10.1007/s11042-020-09067-2.