DOI 10.1007/s10598-023-09568-3
Computational Mathematics and Modeling, Vol. 33, No. 2, April, 2022

HARDNESS ANALYSIS OF X-RAY IMAGES FOR NEURAL-NETWORK TUBERCULOSIS DIAGNOSIS

Ya. A. Pchelintsev,[1,2] **A. V. Khvostikov,**[1,3] **A. S. Krylov,**[1,4]
L. E. Parolina,[5] **N. A. Nikoforova,**[6,7] **L. P. Shepeleva,**[6,8]
E. S. Prokop'ev,[6,9] **M. Farias,**[10] **and Ding Yong**[11]

UDC 519.6+004.891.3

We consider the automatic hardness determination of a chest X-ray image and the effect of pre-filtering of the training and validation samples on the performance of the classification algorithm of tuberculosis diagnosis from chest X-rays. Convolutional neural networks are used in automatic hardness determination and tuberculosis diagnosis. The results of the present study are compared with those from different datasets, including datasets pruned by image hardness criteria.

Keywords: chest X-ray images, tuberculosis diagnosis, quality control, convolutional neural networks, radiograph hardness.

Introduction

Analysis and preprocessing of input data is a topical issue for the application of deep learning methods in medical diagnosis. It is necessary to control the correspondence of input information in the trained (training) deep learning method. This, for instance, is required when controlling for the presence of adversarial attacks on the input data [1].

X-ray hardness is an important factor in radiology and, in particular, tuberculosis diagnosis, as it directly affects the informativeness of the image [2, 3]. Assuming correct contrast of the X-ray image, its hardness can be determined visually by counting the number of upper thoracic vertebrae clearly visible on the X-ray: 3–4 visible vertebrae is the optimal hardness level, a smaller or greater number indicates that the X-ray is too soft or too hard [3, 4]. Examples of X-rays with different hardness levels are shown in Fig. 1.

[1] Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Moscow, Russia.

[2] E-mail: yapchelintsev@live.cs.msu.ru.

[3] E-mail: khvostikov@cs.msu.ru.

[4] E-mail: kryl@cs.msu.ru.

[5] National Medical Research Center of Phthisiopulmonology and Infectious Diseases, Ministry of Health, Moscow, Russia; e-mail: parolina@nmrc.ru.

[6] State Budget Organization of the Sakha (Yakutiya) Republic — N. E. Andreev Scientific-Practical Center "Phthisiatry", Moscow, Russia.

[7] E-mail: nikiada@mail.ru.

[8] E-mail: shepelevalp@mail.ru.

[9] E-mail: Prokopeves@ftiz14.ru.

[10] University of Brasília, Brasilia, Brazil; e-mail: mylene@ene.unb.br.

[11] Zhejiang University, Hangzhou, China; e-mail: dingyong09@zju.edu.cn.

Translated from Prikladnaya Matematika i Informatika, No. 70, 2022, pp. 52–70.

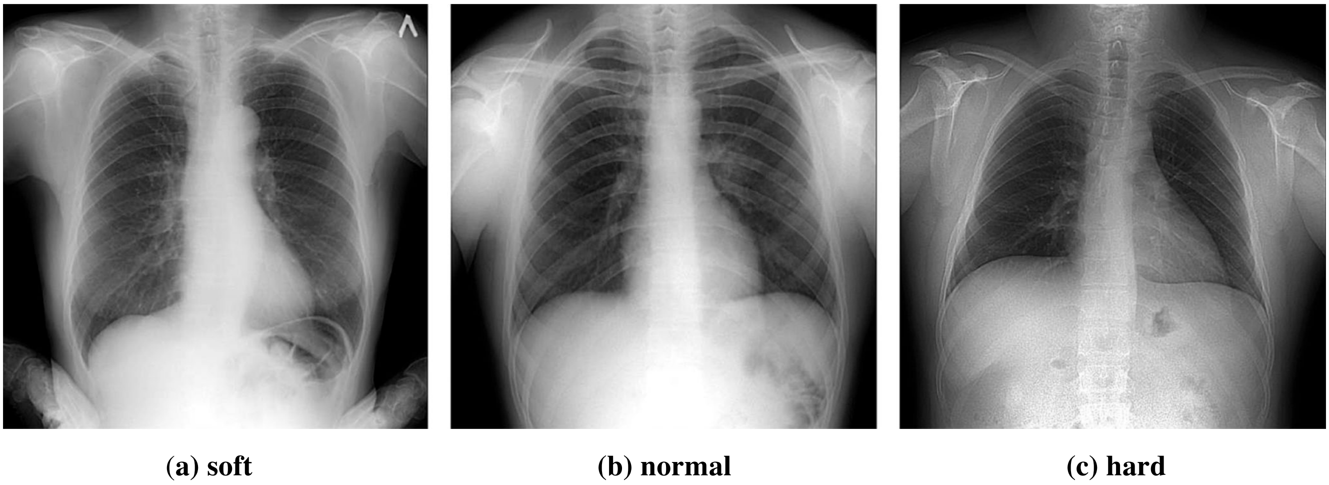

(a) soft **(b) normal** **(c) hard**

Fig. 1. Examples of X-ray images of different hardness levels.

Quality control of chest X-rays based on various parameters is important for a comprehensive analysis of the image and formulation of a correct diagnosis. Automatic determination of the imaging spatial conditions (patient's pose, chest position inside the frame, etc.) is considered in [5, 6, 7]. The effect of image quality on the results of automatic COVID-19 diagnosis is considered in [8].

In our study, the control of input X-ray images is used to verify that the level of radiation is adequate for reliable diagnosis of lung tuberculosis.

This article is a development of [9], where we have shown that optimization of the chest X-ray diagnosis algorithm designed to work with images of close hardness levels, combined with automatic quality control of the X-ray images, ensures better classification accuracy than the algorithm designed to process X-ray images with widely differing hardness levels.

In this article we consider two issues:

(1) automatic hardness determination of chest X-ray images by a neural-network algorithm;

(2) the effect of pre-filtering of the training and validation samples on the classification accuracy for tuberculosis diagnosis from chest X-ray images.

The Data

To develop and test the X-ray hardness determination algorithm, we used a set of 1,298 X-ray images of tuberculosis patients collected in several medical institutions of the Sakha (Yakutiya) Republic. Examples of images from this set are shown in Fig. 2. The set was tagged by a radiologist: for each image, we have the number of clearly visible upper thoracic vertebrae. The distribution of the images by this factor is shown in Fig. 3. In what follows, this dataset is called SakhaTB.

The model training stage for tuberculosis diagnosis is preceded by hardness filtering of the images. Therefore, similarly to the previous study [9], we maintain training, validation, and testing samples of sufficient size for diagnosis by splitting the dataset into two parts, as described below.

The first part of the dataset includes the open-access datasets Montgomery County and Shenzhen [11, 12], which are particularly popular in studies involving X-ray processing of tuberculosis patients [10]. Both datasets

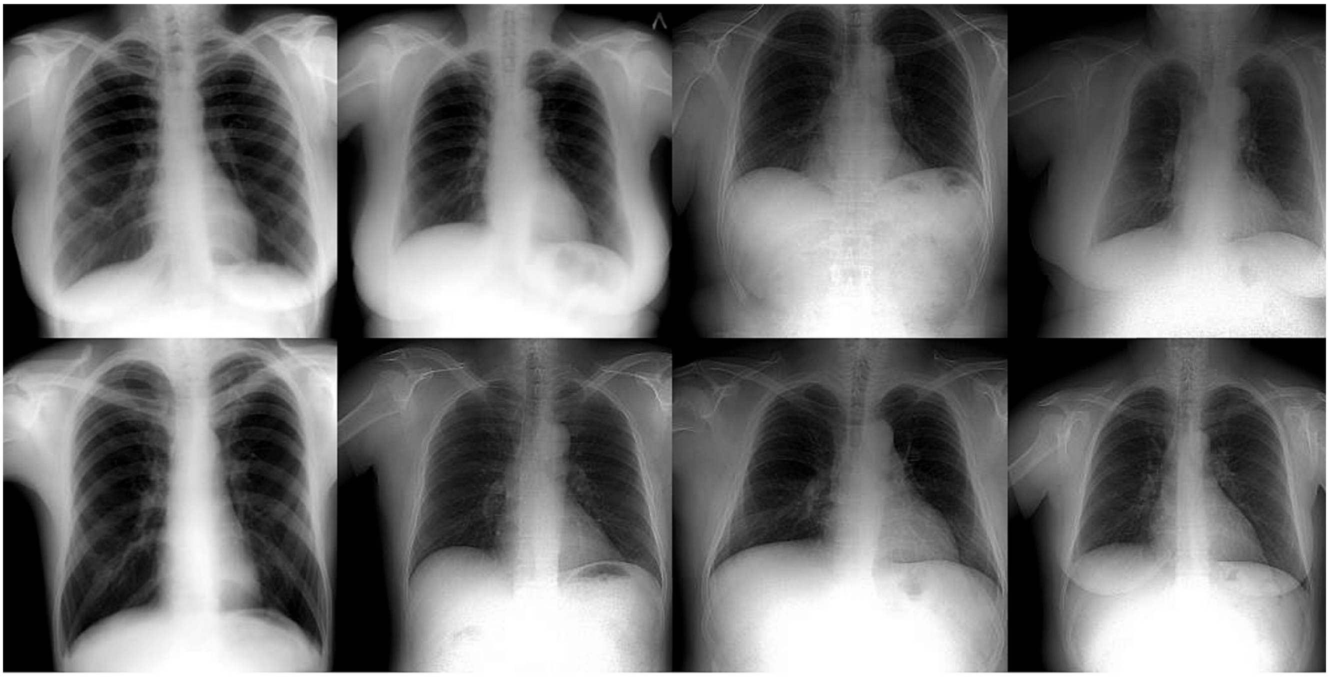


Fig. 2. Examples of X-ray images from the SakhaTB dataset.

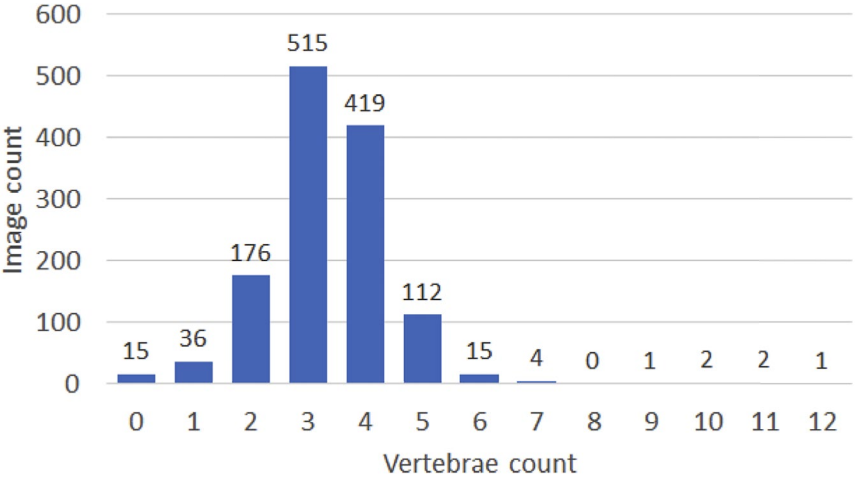
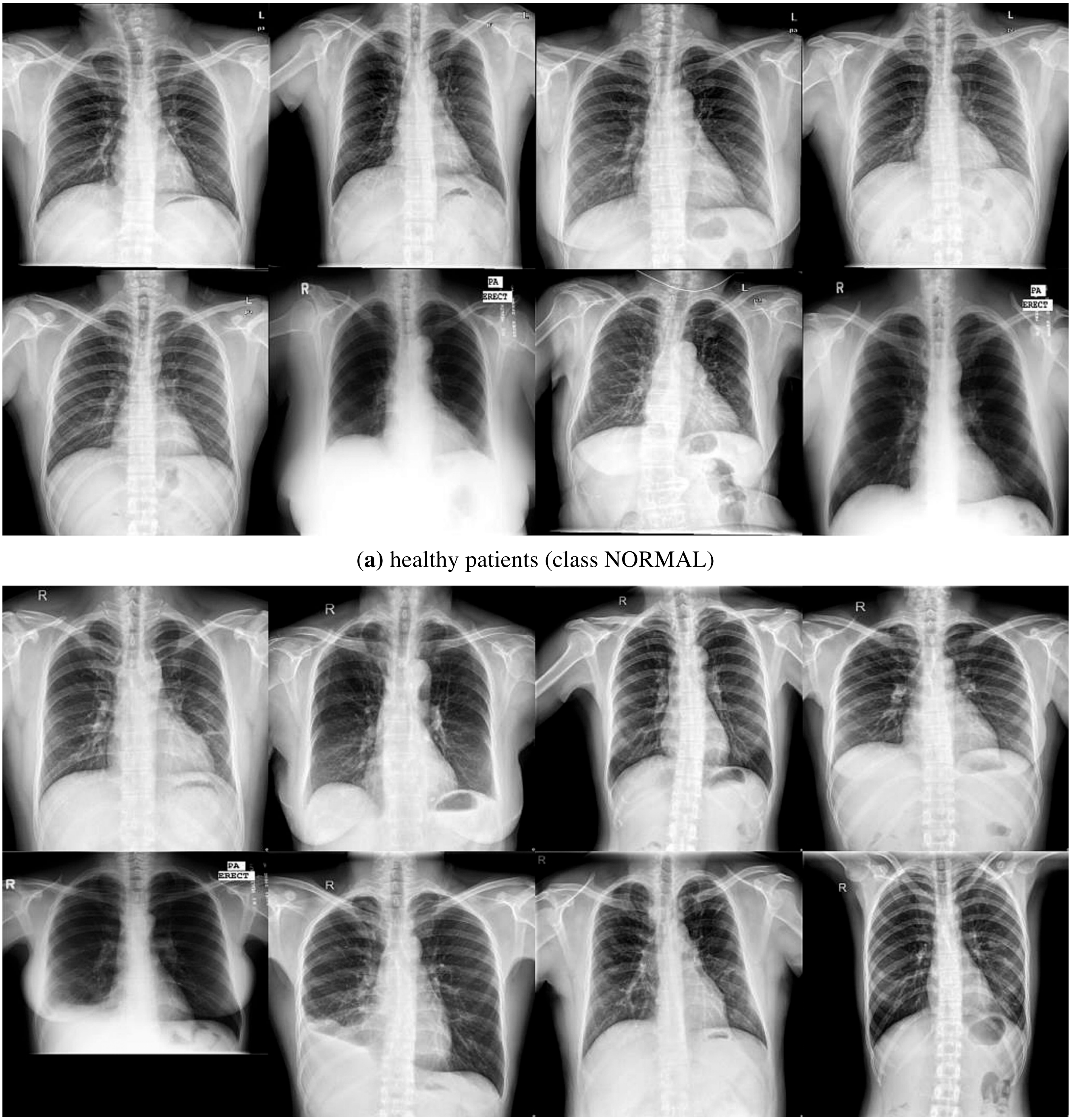


Fig. 3. The distribution of images in the SakhaTB dataset by the count of clearly visible vertebrae.

are available from the NIH National Medical Library and have been collected by the U.S. Department of Health and Social Services in Montgomery County (Maryland) and by the Guangdong Provincial Medical College in People's Hospital No. 3 in Shenzhen (China), respectively. The two datasets contain gray-scale chest X-rays with 8-bit color depth tagged into two classes: healthy individual (NORMAL) and tubercular patient (TB). The image sizes vary and are approximately 3000 × 3000 and 4000 × 4900 pixels. The number of images in each class is presented in Table 1. Since most studies use these two datasets jointly [10], we have also combined them into one dataset (Montgomery-Shenzhen or MC-SZ). Examples of X-ray images from the combined dataset are shown in Fig. 4.

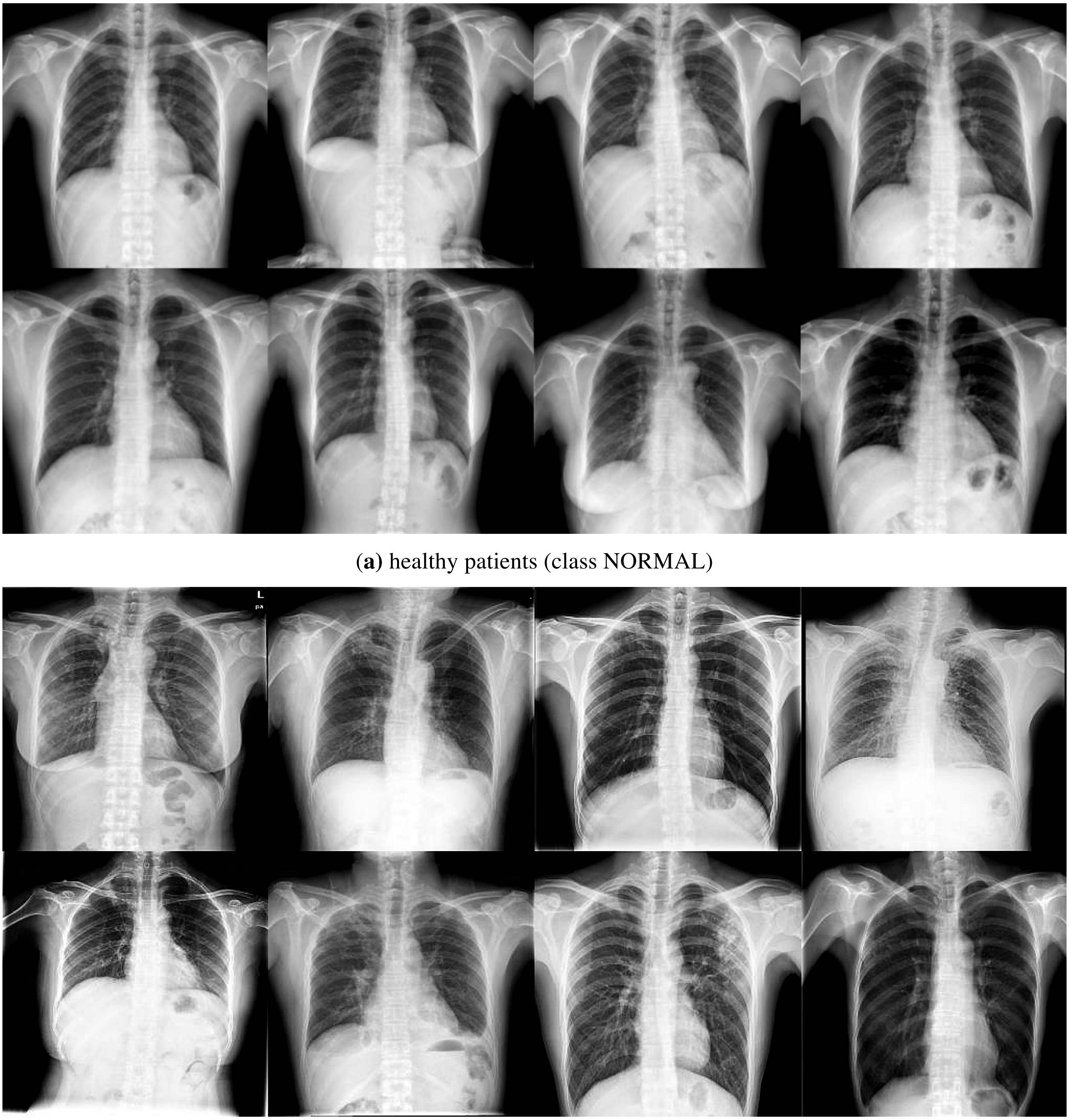


(**a**) healthy patients (class NORMAL)

(**b**) tuberculosis patients (class TB)

Fig. 4. Examples of images from the Montgomery-Shenzhen (MC-SZ) dataset corresponding to different classes.

The second part of the dataset used for tuberculosis diagnosis is a subset of TBX11K [13} prepared in Nankai University (Tianjin, China). TBX11K contains 11,200 gray-scale chest X-rays with 8-bit color depth measuring 512 × 512 pixels. From this number, 8,400 images are tagged to one of three classes (healthy, tuberculosis patients, and patients with other diseases) and show the borders of the diseased lung regions. Table 1 shows the count of tagged images; all were included in the final sample. Examples of X-ray images are shown in Figure 5.



(**a**) healthy patients (class NORMAL)

(**b**) tuberculosis patients (class TB)

Fig. 5. Examples of X-ray images from the TBX11K dataset corresponding to different classes.

A Method of Hardness Determination

The X-ray hardness level is an ordinal quantity, and to preserve the order relations between classes hardness determination is treated as ordinal regression (also known as ordinal classification).

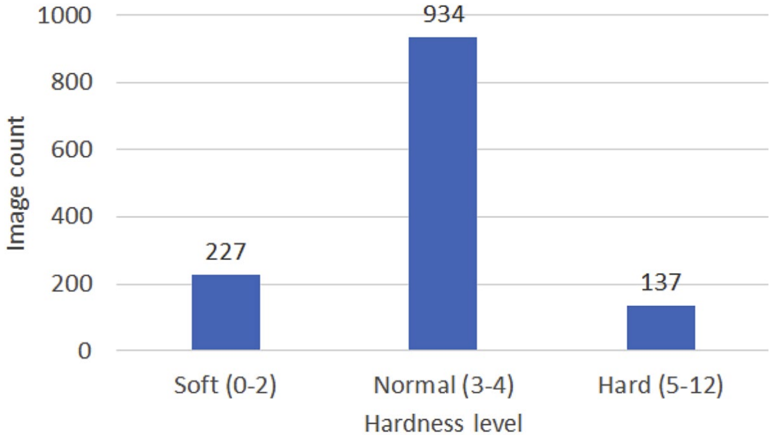


Fig. 6. The distribution of SakhaTB images by hardness level.

Table 1
Size of Datasets Used in This Study

Dataset name	Number of images tagged NORMAL	Number of images tagged TB	Total number of images
Montgomery	80	58	138
Shenzhen	326	336	662
TBX11K	3800	800	4600
Total	**4206**	**1194**	**5400**

We see from Fig. 3 that the set of images is substantially unbalanced and for some hardness sublevels very few cases are available. We have accordingly decided to use the number of clearly visible thoracic vertebrae as identified by a radiologist in order to divide all X-ray images into three hardness groups based on standard medical criteria: soft (fewer than three vertebrae clearly visible), normal (three to four vertebrae clearly visible), and hard (more than four vertebrae) [3, 4]. The count of X-ray images in the three groups is shown in Fig. 6. These three classes were treated as admissible values of the objective variable in ordinal regression.

A neural network was developed for automatic hardness determination of chest X-rays. The input chest X-ray is preprocessed and then assigned a real number from $[0, 1]$ by the neural network; this real number is an internal dimensionless hardness indicator which, after comparison with tunable thresholds, is applied to classify the image in one of the hardness classes. The thresholds are part of the model and are tuned together with the neural network layer weights during training. The advantage of this approach is that the internal hardness indicator can produce a relative ranking of the images even when the current image is substantially different from the training sample so that the class separation thresholds may be incorrectly set for such an image.

Convolutional neural networks of the ResNet family [14], DenseNet [15], and others are widely used in medical image processing and, in particular, in disease diagnosis [10].

In this study, the smallness of the sample suggested using the compact ResNet-18 network with fewer parameters and thus less pronounced tendency to overtrain compared with other networks of the same architectures or representatives of other architectures mentioned above.

We also compared our solution with that produced by a compact network with a newer convolutional neural network architecture EfficientNetV2-S [16] which performs better in image classification problems than any of the networks mentioned above. Its main distinction is the optimization of the functioning of convolutional layers by proportional scaling, change of the order of operations of different dimensions, reduction of convolution kernel size, and omission of "heavy" layers, thus reducing memory use and utilizing the free resources to increase neural network depth and generalizing capacity.

To accelerate training with the aid of ready-made low-level filters in both problems, we used as the initial neural network state the weights of the corresponding model obtained by real-world image classification ImageNet-1K [17]. The last layer was replaced with a fully-connected layer with one output and two tunable threshold parameters in the threshold ordinal regression model [18].

Although the final criterion for determining the X-ray hardness level is the number of clearly visible upper thoracic vertebrae [3, 4], the determination of the contrast level in the preliminary stage requires considering the visibility of other chest regions (for instance, the elements of the lung pattern) and organs [4]. We accordingly decided not to restrict the analysis to the chest but to examine the X-ray image as a whole.

The preprocessing stage includes the following steps:

1. Automatic image contrasting:

$$h(x) = 255 * \frac{x - p_{0.5}}{p_{99.5} - p_{0.5}},$$

where $p_{0.5}(x)$ and $p_{99.5}(x)$ are 0.5% and 99.5% percentiles of the pixel intensities in the image;

2. Automatic gamma-transformation of pixel intensities:

$$\begin{cases} g(x) = 255\left(\frac{x}{255}\right)^{\gamma}, \\ \gamma = \log_{\mu} 0.5, \end{cases}$$

where x is the pixel intensity in the input image, μ the mean intensity of the entire image.

3. Downsizing the image to the input resolution of the neural network (512 × 512 pixels for ResNet-18 and 384 × 384 pixels for EfficientNetV2-S).

4. Optional: global or local (CLAHE [19]) histogram equalization. The side of the window (in pixels) used for local histogram equalization is $\frac{1}{2^n}$ of the image side, where $n \in \mathbb{N}$ is the method parameter. The effect of this step and the window size on the performance of the algorithm will be demonstrated below.

As the loss function we took the all-threshold ordinal regression loss function defined as the sum of terms whose number depends on the number of classes [18]:

$$\begin{cases} L(z, y) = \sum_{k=1}^{K-1} f\left(s(k, y)(\theta_k - z)\right), \\ s(k, y) = \begin{cases} -1, & k < y, \\ +1, & k \geq y, \end{cases} \end{cases}$$

where z is the neural network output (the nondimensional hardness indicator) with values form $[0, 1]$ (the closer to 1, the harder the image), y is the true class of the image corresponding to this output, K is the total number of classes, $\theta_1 < \theta_2 < \ldots < \theta_{K-1}$ are the thresholds partitioning the real axis into K parts, and $f(x)$, is the binary classification loss function used as the base. For the binary classification loss function we take the logistic loss function:

$$f(x) = \ln \frac{1}{1+e^{-x}}.$$

The sample remained highly unbalanced even after reducing the number of ordinal-regression classes to three. We accordingly weighted the loss function for each case with weights inversely proportional to the number of images in the corresponding class. As the measure of ordinal-regression performance we took the mean absolute error balanced by classes (macro-averaged MAE, in what follows mMAE) [20].

The base performance estimate was obtained by training a simple model with classification of images into three classes. The last layer was replaced by a fully-connected layer with three outputs. The loss function was defined as the cross-entropy:

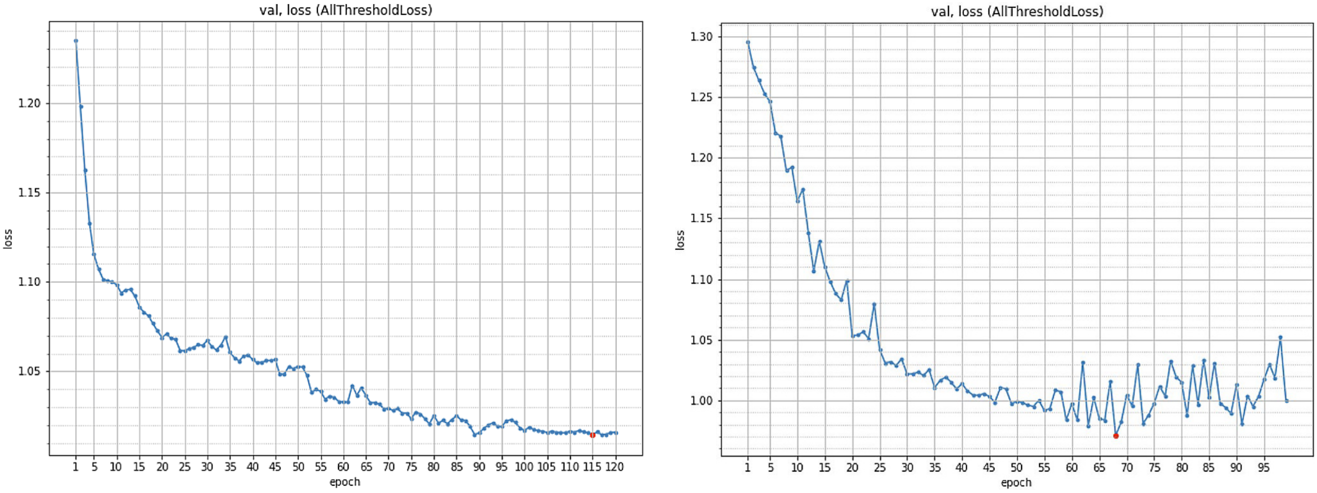
$$\begin{cases} CE(z,y) = \sum_{k=1}^{K} \mathbb{I}[y=k] \cdot \ln\left(softmax(z)_k\right), \\ softmax(z)_k = \dfrac{e^{z_k}}{\sum_{i=1}^{K} e^{z_i}}, \\ I[y=k] = \begin{cases} 1, & y=k, \\ 0, & y \neq k, \end{cases} \end{cases}$$

and the performance measure as the balanced accuracy (in what follows, BalAcc) [21].

The models were trained on a dataset divided into a training, a validation, and a testing sample in 64:16:20 ratio with preliminary random mixing of the images and stratification to preserve the proportions between classes. The training sample underwent random transformations:

- rotations (within 15 degrees in each direction);
- scaling (with a random coefficient from $[0.8, 1.2]$);
- translations (up to 305 of image size along each axis);
- changing brightness and contrast (up to 20% in each direction).

The loss function was optimized by the gradient descent algorithm AdamW [22] with the parameters $lr = 5 \cdot 10^{-6}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\lambda = \dagger 0.01$. The batch size was 64 images for the ResNet-18 model and 16 images for the EfficientNetV2-S model; in both cases, the gradient was accumulated over 8 and 2 iterations respectively (until a “virtual batch” of 128 objects was reached). At the end of each epoch, the model quality was measured on a validation sample; if the loss function on the validation sample had not decreased during 10 epochs, the gradient descent step was reduced by a factor of 5; if there had been no improvement during 31 epochs, training was stopped. Overtraining was controlled by measuring the loss function and the quality



(**a**) ResNet-18

(**b**) EfficientNetV2-S

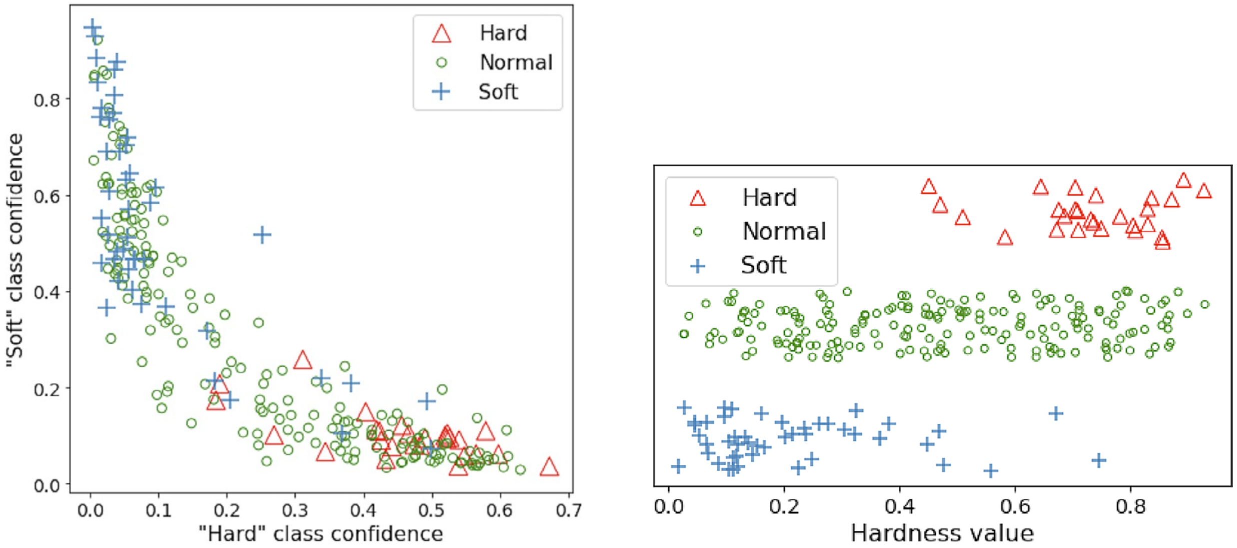
Fig. 7. Plots of the loss function vs. number of epochs on a validation sample for hardness level determination.

Table 2
Quality Measures of the Algorithms Used for Hardness Determination on a Training Sample from SakhaTB

Model	Histogram equalization	BalAcc	mMAE
clf	–	0.563	0.452
ord-eff	–	0.623	0.399
ord	–	0.609	0.452
ord-glob	global	0.609	0.452
ord-clahe2	**1/2 of image side**	**0.636**	**0.398**
ord-clahe4	1/4 of image side	0.610	0.449
ord-clahe8	1/8 of image side	0.593	0.468
ord-clahe16	1/16 of image side	0.600	0.468

measure, but no significant improvement of model performance on the validation sample was observed over time (see Fig. 7; note the slight increase of the model loss function with EfficientNetV2-S, which indicates overtraining), therefore as the final state we took the weights at the end of the last epoch.

The final values of the quality measures obtained on a test samples are presented in Table 2. The models for the solution of ordinal regression models contain “ord” in their names; the model for the classification problem is denoted “clf.” The model “ord-eff” is based on EfficientNetV2-S, all other models are based on ResNet-18. The column “Histogram equalization” is blank if not applied, “global” in case of global equalization, or shows the size of the local equalization window as a fraction of the full image size.



(a) clf (simple classification)

(b) ord-clahe2 (ordinal regression)

Fig. 8. Model predictions versus the true class for an object from SakhaTB.

By balanced accuracy and balanced MAE, the best model was the ordinal regression model with local histogram equalization in the preprocessing stage with a window side measuring 1/2 of the image side. In what follows, we denote this as “ord-clahe2”. Further image analysis is carried out using this model.

Figure 8 plots the probabilities of the classes Hard and Soft predicted by the clf model for the simple classification problem, and also the dimensionless hardness indicator predicted by the ord-clahe2 model for ordinal regression, versus the true hardness value for test-sample objects. The separation of the classes is far from ideal.

Image misclassification was found to have been caused by noise in the dataset due to ambiguity and loose definition of the tagging criteria. This conjecture was also confirmed by the closeness of the training sample evaluation metric to the test sample evaluation metric: balanced accuracy of about 0.70 and 0.67 for clf and ord-clahe2 models respectively.

Analysis of X-ray hardness determination as an ordinal regression in the given model produces, with a certain accuracy, a ranking of the images by hardness based on neural network internal hardness indicator. As a ranking quality measure, we took the Spearman rank correlation coefficient [23] as it is sensitive also to nonlinear correlation. Observations were made for the number of clearly visible vertebrae as reported by the radiologist and the hardness classes. The values of the quality metric are presented in Table 3. Figure 9 shows the distribution of the images from the training sample and from the TBX11K and Montgomery-Shenzhen datasets by the dimensionless hardness indicator predictor of the ord-clahe2 model. Caution should be exercised regarding the separation of images from TBX11K and Montgomery-Shenzhen into hardness classes: the true class thresholds may substantially differ from the network-produced thresholds because the images used may visually diverge from the training images.

We evaluated the performance of the ord-clahe2 model on the MC-SZ dataset. For this purpose, MC-SZ was tagged similarly to SakhatTB: for each image, we counted the clearly visible upper thoracic vertebrae. Table 4 compares the results produced by the algorithm on MC-SZ and on a SakhaTB test sample. Separate histograms for the classes NORMAL and TB from MC-SZ and TBX11K are shown in Fig. 10. The slight differences in the class histograms match the visual differences between the images of these classes: in both datasets there are more soft images in TB and more hard images in NORMAL (see Figs. 4 and 5).

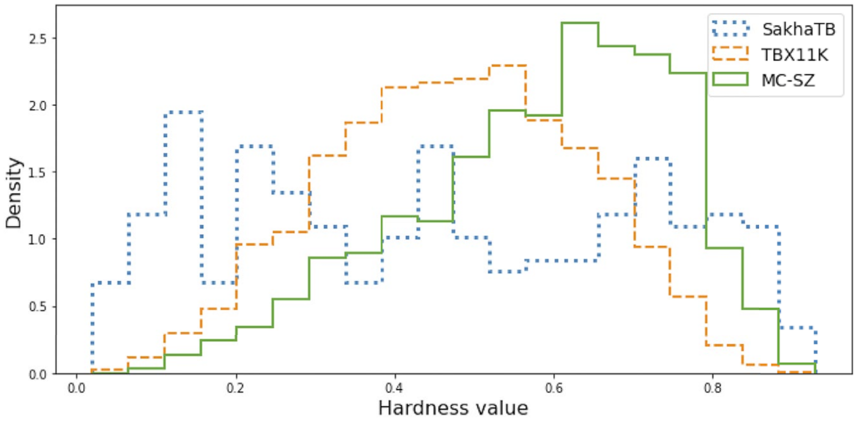


Fig. 9. Distribution of the ord-clahe2 predicted hardness value for images from the three datasets.

Table 3
Ranking Quality Metrics for the SakhaTB Test Sample with Different Hardness Determination Algorithms

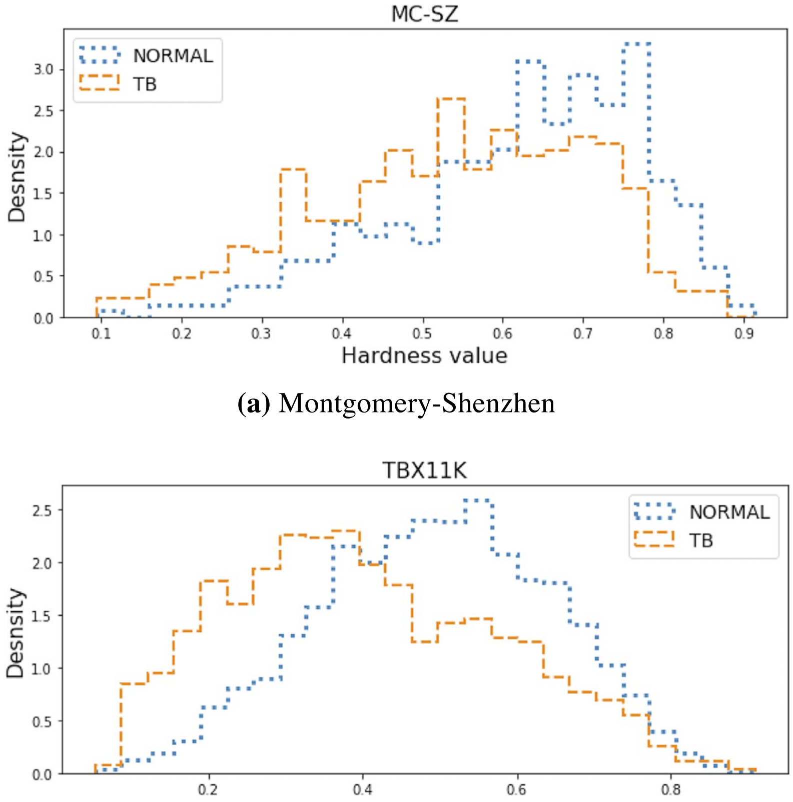
Model	Spearman (vertebrae)	Spearman (hardness)
ord-eff	0.564	0.457
ord	0.576	0.497
ord-glob	0.599	0.514
ord-clahe2	**0.606**	**0.534**
ord-clahe4	0.602	0.519
ord-clahe8	0.588	0.498
ord-clahe16	0.596	0.507

Hardness-Based Neural-Network Tuberculosis Diagnosis

Using the predicted hardness values from the previous stage we removed from the datasets an equal proportion of the hardest and the softest images (i.e., pruning both tails of the distribution). Then the quality of the model trained on the pruned test sample was measured and compared with the quality, on the same test sample, of the model tuned to the unpruned training sample.

Given the visual differences of all three datasets and the small size of the MC-SZ sample, we decided to prune each of the two datasets separately rather than the pooled dataset; in this way, changes of their proportions in the final sample would not affect the quality. This "cautious" approach is associated with the need to perform correct image contrasting before hardness determination, but this issue falls outside the scope of the present article.

In addition to sample pruning from both tails of the hardness level histogram, we considered the case with the omission of only the hardest images, because soft images may preserve some details of lung tissue which are completely lost in hard images.



(**a**) Montgomery-Shenzhen

(**b**) TBX11K

Fig. 10. Distribution of the ord-clahe2 predicted hardness value for images of each class in the two datasets used for tuberculosis diagnosis.

Table 4
Comparison of the Quality of the ord-clahe2 Model on the SakhaTB and MC-SZ Datasets

Dataset	BalAcc	mMAE	Spearman (vertebrae)	Spearman (hardness)
SakhaTB	0.636	0.398	0.606	0.534
MC-SZ	0.546	0.565	0.325	0.203

The tuberculosis diagnosis algorithms may be described as follows: a chest X-ray is delivered to the input for preprocessing; then the neural network assigns two real numbers from $[0, 1]$ to this image (these numbers are the weights of the classes NORMAL and TB). The sum of weights of each image equals 1. The class with the higher weight is accepted as the algorithm output. The neural-network layer weights are adjusted during training.

For neural networks we used EfficientNetV2-S and ResNet-18 with the last layer replaced with a fully-connected two-output layer. The procedures for the division of the dataset into subsamples and model training with weighted classes for balancing, as well as the initial weights and the preprocessing stages were all similar to those in the preceding section; only histogram equalization was omitted. Cross-entropy was used as the loss function and balanced accuracy as the quality metric.

Table 5
Comparison of the Classification Quality of Models on Full and Pruned Datasets (Balanced Accuracy)

	Removal of hard and soft images			Removal of hard images only		
Share removed	5%	10%	15%	5%	10%	15%
Before	0.958	0.951	0.951	0.962	0.961	0.965
After	0.961	0.962	0.953	0.968	0.966	0.975

Table 6
Comparison of the Classification Quality of Models Trained on Complete and Pruned Datasets (Sensitivity/Specificity)

	Removal of hard and soft images			Removal of hard images only		
Share removed	5%	10%	15%	5%	10%	15%
Before	0.923/0.994	0.909/0.994	0.908/0.995	0.930/0.994	0.927/0.995	0.934/0.996
After	0.933/0.990	0.933/0.991	0.915/0.990	0.943/0.994	0.941/0.992	0.958/0.993

The model with EfficientNetV2-S produced better classification before pruning and it was used in all comparisons. This is probably due to the significantly greater size and quality of the sample. The results are presented in Table 5. We see that the change in quality depends on the extent of image pruning but remains always positive. Comparison of sensitivity and specificity indicators for the class TB is shown in Table 6.

CONCLUSION

We have demonstrated the possible use of neutral networks for hardness level determination of chest X-rays. Although high quality measures could not be attained due to the complexity of the problem, the results are significantly better than random choice: balanced accuracy 0.636, Spearman correlation coefficients 0.606 and 0.534. However, the performance of the proposed algorithm noticeably deteriorates when applied to data from other sources: for MC-SZ the balanced accuracy fell from 0.636 to 0.546, and the ranking quality dropped approximately to one-half. To preserve the algorithm performance on images different from the training sample, we have to apply various procedures, such as training sample enlargement, image contrasting with reduction to a single standard, solution of the cross-dataset generalization problem so important in medicine [24]. Application of more rigorous image tagging criteria will probably improve the stability of the method, and the use of contrastive loss and triplet loss as additional loss functions will improve the quality of image ordering relative to one another.

However, even our imperfect model of hardness determination improves the performance of the tuberculosis diagnosis algorithm if the images undergo preliminary filtering before classifier training and generation of predictions. Smaller hardness variability and better data homogeneity substantially improves the detection accuracy of tuberculosis patients ta a cost of a small reduction in specificity: the greatest absolute and relative sensitivity gain for the class TB was observed with removal of 10% of the hardest and 10% of the softest images (from 0.909 to 0.933) and with removal of 15% of the hardest images (from 0.934 to 0.958). In the second instance, we even attained the highest sensitivity (0.958).

Research supported by grants from the Russian Foundation for Basic Research (RFFI 19-5780014) and BRICS (BRICS2019-394).

REFERENCES

1. S. G. Finlayson et al., "Adversarial attacks on medical machine learning," *Science*, **363**, No. 6433, 1287–1289 (2019).
2. G. A. Chuiko and V. M. Tsvetkov, "Effects of X-ray hardness on fluorogram informativeness," *Biomedical Engineering*, **16**, No. 4, 117–119 (1982).
3. L. A. Timofeeva, T. N. Aleshina, and A. V. Bykova, *Main X-ray Syndromes of Lung-Tissue Pathology: a Textbook*, Izd. Chuvash. Univ., Cheboksary (2013).
4. A. U. Sidorov, A. A. Shcherbatykh, and L. N. Pokrovskaya, *Methodology of Radiograph Analysis: a Textbook*, IGMU, Irkutsk (2012).
5. K. Nousiainen et al., "Automating chest radiograph imaging quality control," *Physica Medica*, **83**, 138–145 (2021).
6. J. von Berg et al., "Robust chest x-ray quality assessment using convolutional neural networks and atlas regularization," *Medical Imaging 2020: Image Processing*, *SPIE*, **11313**, 391–398 (2020).
7. J. I. A. Xiao-Qian et al., "Application value of convolutional neural network in quality control of direct digital chest X-ray images," *Xi'an Jiao Tong da Xue Xue Bao*. *Yi Xue Ban*, No. 5, 784 (2019).
8. R. Sadre et al., "Validating deep learning inference during chest X-ray classification for COVID-19 screening," *Scientific Reports*, **11**, No. 1, 1–10 (2021).
9. A. A. Dovganich, A. V. Khvostikov, Y. A. Pchelintsev et al., "Automatic out-of-distribution detection methods for improving the deep learning classification of pulmonary X-ray images," *Journal of Image and Graphics*, **10**, No. 2, (2022).
10. M. Oloko-Oba and S. Viriri, "A systematic review of deep learning techniques for tuberculosis detection from chest radiograph," *Frontiers in Medicine*, **9** (2022).
11. S. Jaeger et al., "Automatic tuberculosis screening using chest radiographs," *IEEE Transactions on Medical Imaging*, **33**, No. 2, 233–245 (2013).
12. S. Candemir et al., "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," *IEEE Transactions on Medical Imaging*, **33**, No. 2, 577–590 (2013).
13. Y. Liu et al., "Rethinking computer-aided tuberculosis diagnosis," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2646–2655 (2020).
14. K. He et al., "Deep residual learning for image recognition," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).
15. G. Huang et al., "Densely connected convolutional networks," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2017), pp. 4700–4708.
16. M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," *International Conference on Machine Learning,* PMLR (2021), pp. 10096–10106.
17. O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, **115**, No. 3, 211–252 (2015).
18. J. D. M. Rennie and N. Srebro, "Loss functions for preference levels: Regression with discrete ordered labels," *Proceedings of the IJCAI Multidisciplinary Workshop on Advances in Preference Handling*, vol. 1, AAAI Press, Menlo Park, CA (2005).
19. S. M. Pizer et al., "Adaptive histogram equalization and its variations," *Computer Vision, Graphics, and Image Processing*, **39**, No. 3, 355–368 (1987).
20. S. Baccianella, A. Esuli, and F. Sebastiani, "Evaluation measures for ordinal regression," *2009 Ninth International Conference on Intelligent Systems Design and Applications*, IEEE (2009), pp. 283–287.
21. K. H. Brodersen et al., "The balanced accuracy and its posterior distribution," *2010 20th International Conference on Pattern Recognition*, IEEE (2010), pp. 3121–3124.
22. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," arXiv preprint arXiv:1711.05101 (2017).
23. D. Zwillinger and S. Kokoska, *CRC Standard Probability and Statistics Tables and Formulae,* Chapman & Hall, New York (2000).
24. V. Thambawita et al., "An extensive study on cross-dataset bias and evaluation metrics interpretation for machine learning applied to gastrointestinal tract abnormality classification," *ACM Transactions on Computing for Healthcare*, **1**, No. 3, 1–29 (2020).