International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-024-00584-z

REGULAR PAPER

Comparison of thresholds for a convolutional neural network classifying medical images

Oona Rainio[1] · Jonne Tamminen[1] · Mikko S. Venäläinen[2,3] · Joonas Liedes[1] · Juhani Knuuti[1] · Jukka Kemppainen[1,4] · Riku Klén[1]

Received: 8 November 2023 / Accepted: 3 June 2024

Abstract
Our aim is to compare different thresholds for a convolutional neural network (CNN) designed for binary classification of medical images. We consider six different thresholds, including the default threshold of 0.5, Youden's threshold, the point on the ROC curve closest to the point (0,1), the threshold of equal sensitivity and specificity, and two sensitivity-weighted thresholds. We test these thresholds on the predictions of a CNN with InceptionV3 architecture computed from five datasets consisting of medical images of different modalities related to either cancer or lung infections. The classifications of each threshold are evaluated by considering their accuracy, sensitivity, specificity, F1 score, and net benefit. According to our results, the best thresholds are Youden's threshold, the point on the ROC curve closest to the point (0,1), and the threshold of equal sensitivity and specificity, all of which work significantly better than the default threshold in terms of accuracy and F1 score. If higher values of sensitivity are desired, one of the two sensitivity-weighted could be of interest.

Keywords Classification · Convolutional neural network · Medical imaging · Thresholds

1 Introduction

During the past decade, the amount of collected data has increased and the use of artificial intelligence in analysis of large datasets has become vital. Typically, a large dataset and machine learning are used to build a predictive model, which can be used to predict the outcome for new samples. Especially in biomedical field, many of the predictive models are used for binary outcome but, even if the outcome is expected to be binary, the models may produce a numeric value such as probability of the outcome and converting this value into binary prediction is not a trivial problem. In particular, this issue is present when using one very popular deep learning technique called a convolutional neural network (CNN) for classifying images between two categories.

In the literature, several methods for identifying the optimal decision threshold have been proposed [5, 6, 22, 24] but there is very little systematic comparison between these thresholds, at least for machine learning applications. Consequently, the output of a CNN is very often converted with either the default threshold of 0.5 or Youden's threshold [24]. CNNs are typically trained with annotated data, meaning that they can be considered an example of supervised learning, and their predictions of the test set can similarly be eval-

✉ Oona Rainio
ormrai@utu.fi

Jonne Tamminen
jonne.j.tamminen@utu.fi

Mikko S. Venäläinen
mikko.venalainen@utu.fi

Joonas Liedes
joolie@utu.fi

Juhani Knuuti
Juhani.Knuuti@tyks.fi

Jukka Kemppainen
Jukka.Kemppainen@tyks.fi

Riku Klén
riku.klen@utu.fi

[1] Turku PET Centre, University of Turku and Turku University Hospital, Turku, Finland

[2] Turku Bioscience Centre, University of Turku and Åbo Akademi, Turku, Finland

[3] Department of Medical Physics, Turku University Hospital, Turku, Finland

[4] Department of Clinical Physiology and Nuclear medicine, Turku University Hospital, Turku, Finland

Published online: 18 June 2024

uated by comparing these predictions to their ground-truth values. Since many evaluation metrics interesting in diagnostics, such as accuracy, sensitivity, specificity, and F1 score, require converting numerical predictions into binary labels [17], it is important to study the impact different thresholds might have to the final results.

Furthermore, it must be noted that there are different measures and criteria for assessing the performance of the CNNs. Many potential methods aim to the highest possible accuracy and give equal weights to both sensitivity and specificity but this assumption is not always justified when different ethical and financial factors are also considered. For example, in the case of a fatal illness that would be avoidable with a relatively inexpensive test, giving more weight on sensitivity might be more meaningful to avoid all possible occurrences of the disease. Alternative methods with different weighting factors for sensitivity and specificity have also been discussed [12, 13, 19], but they have not gained widespread attention and their applicability have been demonstrated only in a limited number of datasets. Given it is natural that sensitivity-weighted methods yield higher sensitivity and lower accuracy and specificity, different evaluation metrics such as F1 score (also known as Dice score) or some measure of net benefit are needed.

In the present study, we compare six different thresholds for converting continuous predictions of a CNN. These thresholds include the default threshold of 0.5, Youden's threshold, the point on the receiver operating characteristic (ROC) curve closest to the point (0,1), the threshold of equal sensitivity and specificity, and two different sensitivity-weighted thresholds. In this comparison, we use five different medical image datasets related to cancer and certain lung infections. The performance of each threshold is by using accuracy, sensitivity, specificity, the F1 score, and the net benefit. Our aim is to give the reader a basic understanding of the existing thresholds, quantify how great differences they can cause in the evaluation metrics, and study the new sensitivity-weighted thresholds. The codes for applying the metrics in practice are also made freely available in several programming languages.

2 Materials and methods

2.1 Software requirements

The experiments of this article were run in Python (version: 3.9.9) [18] with the packages TensorFlow (version: 2.7.0) [1] and Keras (version: 2.7.0) [2].

2.2 Data

Five different datasets are studied. The first of them is private and the other four are from publicly available repositories. In each dataset, there are equally many images considered positive as there are negative images. The latter four datasets are selected so that each image is from a different patient by removing additional images if necessary.

The first data contain 1115 transaxial slices depicting a tumor from chosen from the positron emission tomography (PET) images of 100 different head and neck cancer (HNC) patients diagnosed with either head and neck squamous cell carcinoma, adenocarcinoma, adenoid cystic carcinoma, or parotid cancer, and equally many randomly chosen negative slices from the PET images of 100 patients who were previously diagnosed HNC but had no locoregional recurrences after curative chemoradiotherapy. The patients were imaged with 3T Philips Ingenuity TF PET/magnetic resonance imaging (MRI) scanner (Philips Health Care) or SIGNA PET/MRI with QuantWorks (GE Healthcare) by using 18F-fluorodeoxyglucose as tracer substance in Turku PET Centre in Turku, Finland during years 2014-2022. A singular three-dimensional PET image consist of 32-66 transaxial slices of 512×512 pixels so that each voxel is of size of $4\text{mm} \times 4\text{mm} \times 4\text{mm}$. The positive slices were chosen according to binary masks created with Carimas [16] by a medical doctor. All the participants were at least 18 years of age, consented to research use of their data, and the research from their data was approved by Ethics Committee of the Hospital District of Southwest Finland.

The second dataset has 3000 transaxial images of MRI scans of different patients with or without brain tumor (BT) so that each image shows their brains similarly and the tumor is visible in the images of the positive patients. The last three datasets all consist of chest X-rays of patients diagnosed with some type of lung infection and healthy patients. The positive patients of the third dataset have COVID-19 (CoV), the patients of the fourth data have pneumonia (PNA), and the patients of the fifth data have tuberculosis (TB). The numbers of images in these three datasets are 3000, 3000, and 1400, respectively. More details for the last four datasets can be found from their repositories and cited references: Br35h:: Brain Tumor Detection 2020 [8], COVID-19 Radiography Database [4, 14], Chest X-Ray Images (Pneumonia) [10], and Tuberculosis (TB) Chest X-ray Database [15] (see the links from the data availability statement). The key details of each five datasets are summarized in Table 1 and Fig. 1 contains a few example images of the positive cases.

Table 1 The imaging modality, the imaged area, the anatomic direction of the images, the diagnosis of the positive patients, the number of images, and the reference for the five datasets

Data	Modality	Area	Direction	Diagnosis	Images	Ref
HNC PET	PET	Head and neck	Transaxial	Tumor	2230	[9]
BT MRI	MRI	Brain	Transaxial	Tumor	3000	[8]
CoV X-ray	X-ray	Chest	Coronal	COVID-19	3000	[4, 14]
PNA X-ray	X-ray	Chest	Coronal	Pneumonia	3000	[10]
TB X-ray	X-ray	Chest	Coronal	Tuberculosis	1400	[15]

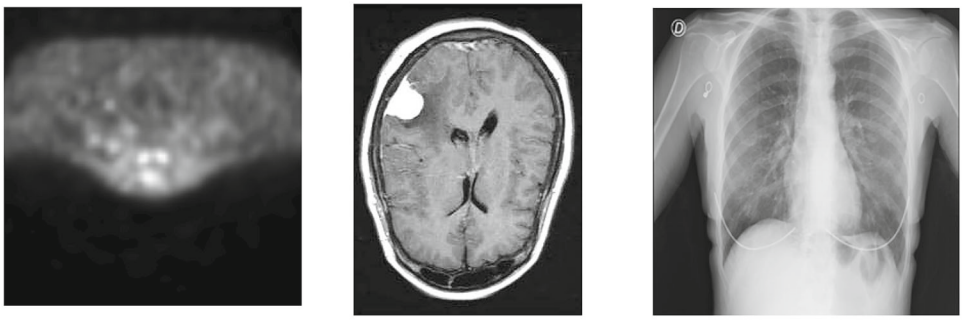

Fig. 1 Examples of the positive images from the first three datasets before pre-procesing: (A) A PET image slice showing the neck and shoulders area of patient with a hypopharynx squamous cell carcinoma, (B) an MRI from a tumorous brain, and (C) a chest X-ray of a patient diagnosed with COVID-19

2.3 Pre-processing

The grayscale images are converted to the size of $128 \times 128 \times 3$ matrices with pixel values as integers varying on the usual interval [0,255].

2.4 Cross-validation

Fivefold cross-validation is used to create the training and test sets. For all the datasets, the data were divided patient-wise into five possible test sets so that the size of a test set is always exactly 20% of the total data. This division was also done so that there are equally many positive and negative images within both the training set and the test set.

2.5 Convolutional neural network

The CNN used in this study is InceptionV3 introduced by Szegedy et al. [21], which is a state-of-the-art classification CNN readily available in the Keras library. The aim of InceptionV3 was to minimize the computational cost without loosing the generalization ability of the deeper network [3]. In the architecture, this was achieved by using smaller asymmetric filters and bottleneck convolutions to reduce data dimensions [3]. The exact architecture is presented in Table 1 of [21]. The InceptionV3 CNN is loaded from Keras without the pre-trained weights and instead only trained on our datasets. The binary cross-entropy is used as the loss function and optimizer is the stochastic gradient descent with a learning rate of 0.001. During the training, the CNN uses 30% of the training data for validation. Based on initial tests on converge, the number of epochs is 10 for all the datasets.

2.6 Thresholds

Denote sensitivity (percentage of positive instances classified correctly) and specificity (percentage of negative instances classified correctly) by $sens$ and $spec$, respectively. The ROC curve is the curve obtained by plotting sensitivity against the false positive rate (equal to $1 - spec$). Then we can define the following thresholds:

1.) *Default* threshold of 0.5,
2.) *Youden*'s threshold [24] that is found by maximizing the value of $sens + spec - 1$,
3.) *minROCdist* that is found by minimizing the distance $\sqrt{(1 - sens)^2 + (1 - spec)^2}$ from the point (0,1) to the ROC curve, and
4.) threshold *equsen* of equal sensitivity and specificity found by minimizing the absolute value of their difference.

There are other possible thresholds, such as the threshold of maximum accuracy and Cohen's kappa considered in [7] but, for datasets with equally many positive and negative instances, they become identical to Youden's threshold.

Let us then define two sensitivity-weighted thresholds as follows: For a parameter $c > 0$, let

5.) *sendist* be the threshold minimalizing distance $\sqrt{(1+c-sens)^2+(1-spec)^2}$ from the point (0,1+c) to the ROC curve, and
6.) *sencp* be the threshold maximizing the product $sens(spec+c)$.

The threshold sendist is a sensitivity-weighted version of minROCdist while sencp is a similar modification of the threshold of the concordance probability method [11], which is based on maximizing the product of sensitivity and specificity. As we do not know which value of the parameter c performs the best and want compare these two sensitivity-weighted thresholds with each other rather obtain some specific value for sensitivity, we fix $c = 0.5$ for both thresholds for the experiments.

2.7 Evaluation metrics

In addition to sensitivity and specificity, we consider accuracy (percentage of instances classified correctly) and the F1 score (harmonic mean of precision and sensitivity) to evaluate how well the thresholds work. Furthermore, we also use the following evaluation metric: For a given probability threshold p, the net benefit is defined by

$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} - \frac{p}{1-p}\cdot\frac{\mathrm{FP}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}, \tag{2.1}$$

where TP, FP, TN, and FN mean the numbers of true positive, false positive, true negative, and false negative predictions. Since there are equally many positive and negative images in our datasets, we choose here $p = 1/2$, which means that the net benefit is the difference between TP and FP predictions divided by the total number of the predictions. Unlike the earlier evaluation metrics, whose possible values range on the interval [0,1], the potential range of net benefit is now $[-0.5,0.5]$.

2.8 Structure of the experiment

For all five datasets, the CNN is initialized, trained with training data, and used to predict the images of the test set for five times for each different test set of the fivefold cross-validation. During each iteration round, the thresholds are chosen according to training data, the output of the test set is converted according to the found thresholds, and the values of the five evaluation metrics are computed. The values of the evaluation metrics are summarized with their mean and standard deviation over the five data divisions. The Wilcoxon signed-rank test is then used to estimate whether the differences in the evaluation metrics for different thresholds are statistically significant or not [17].

3 Results

According to Wilcoxon signed-rank tests, different thresholds result in statistically significant differences in the values of the evaluation metrics. The mean values and standard deviation of accuracy, sensitivity, specificity, and F1 score are presented in Table 2 for different thresholds computed from the five datasets. Notably, the default threshold has always highest specificity but nearly always the lowest mean for accuracy, sensitivity, and F1 score. By computing the mean values of the mean accuracies in Table 2, it can actually be seen that Youden's threshold, minROCdist, and equsen have the highest accuracy on average and both the sensitivity-weighted thresholds outperform the default threshold. Also, the Youden's threshold, minROCdist, and equsen give relatively similar mean values for sensitivity and specificity, while the default threshold results in higher specificity than sensitivity.

From Table 2, we also see that sendist gives always the highest mean value for sensitivity. However, it performs quite poorly in terms of specificity, even though its mean accuracy is close to that of the default threshold. While sencp has lower sensitivity than equsen, this threshold works very well when measured by accuracy and F1 score.

Table 3 contains the mean values and standard deviation of net benefit. Surprisingly, we see that the highest net benefit is often obtained with the default threshold. This suggest that the TN observations have higher weight in the definition of the net benefit than in accuracy or F1 score. Still, the differences in the netf benefit between the default threshold and Youden's threshold, minROCdist, or equsen are quite small. The sensitivity-weighted sendist threshold has always the lowest net benefit.

4 Discussion

According to our results, Youden's threshold, minROCdist, or equsen is the best threshold if there is no special need for high sensitivity or specificity. The use of the default threshold should be considered carefully because all the three other non-sensitivity-weighted thresholds work significantly better in terms of accuracy and F1 score. However, for certain definitions of the net benefit, the use of the default threshold might be justified. It should be taken into account that our tests also suggest that the default threshold is often very imbalanced in terms of sensitivity and specificity in a way that causes much higher values of specificity even though there would be equally many positive and negative instances in the data.

Out of the two sensitivity-weighted thresholds, sencp is less sensitive to values of the parameter c than sendist. It depends on the requirements of the scientist, which of these

Table 2 The mean and the standard deviation as percents for accuracy (Acc.), sensitivity (Sen.), specificity (Spec.), and F1 score computed by using different thresholds for the predictions from the five datasets. On each row, the highest mean is in bold and the lowest mean in italics

Data	Metric	Default	Youden	minROCdist	equsen	sendist	sencp
HNC	Acc	76.6±3.0	78.1±3.3	**78.4±3.0**	78.3±3.0	*76.4±2.7*	77.8±2.5
PET	Sen	*67.6±5.2*	77.6±6.3	78.2±5.0	78.4±4.3	**85.7±4.2**	82.7±4.7
	Spec	**85.5±2.2**	78.6±3.7	78.5±2.7	78.2±2.6	*67.2±2.5*	72.8±2.7
	F1	*74.2±3.9*	77.9±3.9	78.3±3.3	78.3±3.2	78.4±2.7	**78.8±2.7**
BT	Acc	*69.3±4.9*	71.6±5.6	71.6±5.8	**71.8±5.8**	70.7±5.2	71.4±5.2
MRI	Sen	*54.3±6.3*	72.8±8.4	71.8±7.6	71.8±7.3	**82.6±6.5**	80.2±6.6
	Spec	**84.3±4.2**	70.4±5.0	71.4±4.6	71.7±4.7	*58.8±5.1*	62.6±5.1
	F1	*63.8±6.2*	71.8±6.3	71.5±6.3	71.7±6.2	**73.8±4.9**	73.7±5.1
CoV	Acc	*76.9±4.6*	78.9±4.7	**79.0±4.7**	**79.0±4.7**	77.4±3.8	78.1±4.2
X-ray	Sen	*65.5±5.6*	80.3±5.4	79.2±3.9	79.3±4.3	**88.5±2.7**	85.8±4.4
	Spec	**88.3±4.3**	77.5±6.4	78.8±5.9	78.7±5.4	*66.2±5.7*	70.5±5.1
	F1	*73.9±5.5*	79.2±4.7	79.1±4.4	79.1±4.6	**79.7±3.2**	**79.7±3.8**
PNA	Acc	86.2±3.7	**86.6±3.8**	86.4±4.0	86.3±3.9	*84.3±3.9*	86.1±3.9
X-ray	Sen	*78.4±5.2*	84.0±4.1	85.5±3.7	86.7±3.2	**90.6±3.0**	87.5±3.5
	Spec	**93.9±2.7**	89.2±4.6	87.3±4.9	85.9±5.3	*78.0±6.3*	84.8±5.1
	F1	*84.9±4.3*	86.2±3.9	86.3±3.9	**86.4±3.7**	85.3±3.4	86.3±3.7
TB	Acc	*65.9±11.4*	69.3±11.9	69.3±11.9	69.1±11.8	**69.4±10.3**	69.5±11.5
X-ray	Sen	*50.1±11.7*	73.6±13.7	70.9±14.0	68.4±11.9	**83.2±10.5**	79.7±13.8
	Spec	**81.8±11.5**	65.0±11.1	67.7±10.2	69.8±11.9	*55.7±11.0*	59.2±10.4
	F1	*59.5±13.4*	70.4±12.1	69.5±12.5	68.9±11.9	**73.1±9.1**	72.1±11.4

Table 3 The mean and standard deviation of the net benefit (possible range [−0.5,0.5]) computed by using different thresholds for the predictions from the five datasets. On each row, the highest mean is in bold and the lowest mean in italics

Data	Default	Youden	minROCdist	equsen	sendist	sencp
HNC PET	**0.291±0.028**	0.282±0.029	0.284±0.026	0.283±0.027	*0.242±0.023*	0.264±0.021
BT MRI	**0.227±0.054**	0.213±0.053	0.215±0.055	0.217±0.055	*0.185±0.045*	0.196±0.047
CoV X-ray	**0.304±0.050**	0.285±0.050	0.290±0.050	0.290±0.049	*0.247±0.038*	0.262±0.040
PNA X-ray	**0.392±0.036**	0.376±0.041	0.368±0.043	0.362±0.043	*0.324±0.043*	0.357±0.042
TB X-ray	0.189±0.133	0.184±0.115	0.188±0.116	**0.193±0.119**	*0.171±0.092*	0.175±0.105

thresholds should be used and how the value of c should be fixed. Our tests suggest that the choice $c = 0.5$ works very well for the sencp threshold, if high values of F1 scores are desired. If the aim is to reach some known level of sensitivity, one alternative method would be choose directly the threshold that gives this required value of sensitivity for the training data. Furthermore, if there is no special need for a single optimal decision threshold, the diagnostic models can also be compared via a ROC curve analysis (see, for instance, [6, 20, 23]).

Another option for obtaining a specific value of sensitivity would be using weights in the training of the CNN. In this way, the user can specify how much difference is there between the importance of classifying a positive instance correctly and classifying a positive instance correctly. However, the issue of this method is that the weights of sensitivity cannot be replaced after the training without re-training. This is one of the advantage of using different thresholds: As long as the initial predictions of both training and test sets are saved, the classifications can be very quickly computed with any given threshold.

One of the evaluation metrics used here, the net benefit, has a parameter called probability threshold (see formula (2.1)). A typical use of net benefit involves clinical knowledge of trade-off between cases and controls, which is related to the probability threshold. If such trade-off is not known, then net benefit is often plotted for a range of suitable probability threshold values. We wanted to obtain a single net benefit value and thus in our analysis we fixed the probability threshold to be $1/2$ based on the ratio of number of positive outcomes to the sample size of the dataset. This normalization allows the comparison of net benefit between different datasets but testing the effect of varying probability thresh-

old could be also considered, which might offer an interesting topic for further study.

There are also several other topics that could be considered in future research. We only studied here classification of two-dimensional images but, in the medical field, CNNs are also commonly used for both three-dimensional images and for image segmentation instead of classification. Additionally, also thresholds for deep learning models other than CNNs, such as vision transformers, could be studied.

5 Conclusion

Different thresholds were compared for a CNN with InceptionV3 architecture used for binary classification of different datasets of medical images. According to our results, the best thresholds for a CNN classifying medical images are Youden's threshold, the point on the ROC curve closest to the point (0,1), and the threshold of equal sensitivity and specificity. In particular, in terms of accuracy and F1 score, they outperform the default threshold of 0.5 in a statistically significant way. We also introduced two sensitivity-weighted thresholds that might be useful in certain applications.

Author Contributions OR wrote the final manuscript, performed the tests, and prepared all the figures and tables. JT and MV wrote an early version of the manuscript and performed initial tests. JL pre-processed data. JK, JK, and RK supervised the project.

Funding Open Access funding provided by University of Turku (including Turku University Central Hospital). The first author was financially supported by the Finnish Culture Foundation.

Data availability The first dataset is private due to ethical restrictions. The other four datasets are publicly available online. They include Br35h:: Brain Tumor Detection 2020 [8] https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection, COVID-19 Radiography Database [4, 14] https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database, Chest X-Ray Images (Pneumonia) [10] https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia, and Tuberculosis (TB) Chest X-ray Database [15] https://www.kaggle.com/datasets/tawsifurrahman/tuberculosis-tb-chest-xray-dataset.

Code availability Available at https://github.com/rklen/threshold_comparison_for_a_CNN

Declarations

Conflict of interest On the behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Zheng, X.: TensorFlow: large-scale machine learning on heterogeneous systems (2015)
2. Chollet, F. et al.: Keras. GitHub (2015)
3. Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M., Farhan, Laith: Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J. Big Data **8**, 53 (2021)
4. Chowdhury, M.E.H., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M.A., Mahbub, Z.B., Islam, K.R., Khan, M.S., Iqbal, A., Al-Emadi, N., Reaz, M.B.I., Islam, M.T.: Can AI help in screening Viral and COVID-19 pneumonia? IEEE Access **8**, 132665–132676 (2020)
5. Coffin, M., Sukhatme, S.: Receiver operating characteristic studies and measurement errors. Biometrics **53**, 823 (1997)
6. Faraggi, D.: Adjusting receiver operating characteristic curves and related indices for covariates. J. R. Stat. Soc. Ser. D Stat. **52**, 179–192 (2003)
7. Freeman, E.A., Moisen, G.G.: A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. Ecol. Modell. **217**(1–2), 48–58 (2008)
8. Hamada, A.: Br35h:: Brain tumor detection 2020, version 12, accessed on Feb 24th, 2023. https://www.kaggle.com/ahmedhamada0/brain-tumor-detection (2020)
9. Hellström, H., Liedes, J., Rainio, O., Malaspina, S., Kemppainen, J., Klén, R.: Classification of head and neck cancer from PET images using convolutional neural networks. Sci. Rep. **13**, 10528 (2023)
10. Kermany, D.S., Goldbaum, M., Cai, W., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell **172**(5), 1122-1131.e9 (2018)
11. Liu, X.: Classification accuracy and cut point selection. Stat. Med. **31**, 2676–2686 (2012)
12. Li, D.-L., Shen, F., Yin, Y., et al.: Weighted youden index and its two-independent-sample comparison based on weighted sensitivity and specificity. Chinese Med. J. **126**, 1150–1154 (2013)
13. Perkins, N.J., Schisterman, E.F.: The inconsistency of 'optimal' cutpoints obtained using two criteria based on the receiver operating characteristic curve. Am. J. Epidemiol. **2006**(163), 670–675 (2006)
14. Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Kashem, S.B.A., Islam, M.T., Maadeed, S.A., Zughaier, S.M., Khan, M.S., Chowdhury, M.E.: Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. Comput. Biol. Med., Vol. 132, 104319 (2021)
15. Rahman, T., Khandakar, A., Kadir, M.A., Islam, K.R., Islam, K.F., Mahbub, Z.B., Ayari, M.A., Chowdhury, M.E.H.: Reliable tubercu-

losis detection using chest X-ray with deep learning, segmentation and visualization. IEEE Access **8**, 191586–191601 (2020)
16. Rainio, O., Han, C., Teuho, J., Nesterov, S.V., Oikonen, V., Piirola, S., Laitinen, T., Tättäläinen, M., Knuuti, J., Klén, R.: Carimas: an extensive medical imaging data processing tool for research. J. Digit Imag. **36**, 1885–1893 (2023)
17. Rainio, O., Teuho, J., Klén, R.: Evaluation metrics and statistical tests for machine learning. Sci. Rep. **14**, 6086 (2024)
18. van Rossum, G., Drake, F.L.: Python 3 reference manual. CreateSpace (2009)
19. Rucker, G., Schumacher, M.: Summary ROC curve based on a weighted Youden index for selecting an optimal cutpoint in meta-analysis of diagnostic accuracy. Stat. Med. **29**, 3069–3078 (2010)
20. Schisterman, E.F., Faraggi, D., Reiser, B.: Adjusting the generalized ROC curve for covariates. Stat. Med. **23**, 3319–3331 (2004)
21. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp. 2818–2826 (2016)
22. Tilbury, J.B., Van Eetvelt, P.W.J., Garibaldi, J.M., et al.: Receiver operating characteristic analysis for intelligent medical systems—a new approach for finding confidence intervals. IEEE Trans. Biomed. Eng. **47**, 952–963 (2000)
23. Zhou, X.-H., McClish, D.K., Obuchowski, N.A.: Statistical Methods in Diagnostic Medicine (2009)
24. Youden, W.J.: Index for rating diagnostic tests. Cancer **3**(1), 32–35 (1950)