DR-GPT: a large language model for medical report analysis of diabetic retinopathy patients

Joel Jaskari[1], Jaakko Sahlsten[1], Paula Summanen[2], Jukka Moilanen[2], Erika Lehtola[2], Marjo Aho[3], Elina Säpyskä[3], Kustaa Hietala[4], Kimmo Kaski[1,5*],

1 Department of Computer Science, Aalto University, Espoo, Finland
2 Department of Ophthalmology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland
3 Department of Ophthalmology, Helsinki University Hospital and University of Helsinki, Helsinki, Finland
4 Central Finland Health Care District, Jyväskylä, Finland
5 The Alan Turing Institute, London, United Kingdom

* Corresponding author, kimmo.kaski@aalto.fi

Abstract

Diabetic retinopathy (DR) is a sight-threatening condition caused by diabetes. Screening programmes for DR include eye examinations, where the patient's fundi are photographed, and the findings, including DR severity, are recorded in the medical report. However, statistical analyses based on DR severity require structured labels that calls for laborious manual annotation process if the report format is unstructured. In this work, we propose a large language model DR-GPT for classification of the DR severity from unstructured medical reports. On a clinical set of medical reports, DR-GPT reaches 0.975 quadratic weighted Cohen's kappa using truncated Early Treatment Diabetic Retinopathy Study scale. When DR-GPT annotations for unlabeled data are paired with corresponding fundus images, the additional data improves image classifier performance with statistical significance. Our analysis shows that large language models can be applied for unstructured medical report databases to classify diabetic retinopathy with a variety of applications.

Introduction

Diabetic retinopathy (DR) is a sight-threatening eye disease that develops as a result of diabetes. As a standard practice, ophthalmologists first identify signs of the disease by observation and then classify the disease based on the combination and severity of the visible signs. In Finland, the photographer and physician record the findings and resulting DR classification in the patient's electronic health records, possibly in an unstructured manner depending on the healthcare provider. In order to perform any computational analyses on a patient or population level, the DR classification label is required. However, manual labeling of retrospectively collected medical reports can become prohibitive when dealing with large scale studies involving tens of thousands of patients.

Recently, large language models (LLMs) have gained widespread popularity among the public with the rise of chat-based applications, such as OpenAI's ChatGPT, Meta's Llama, and Google's Bard. In the medical domain, LLMs are being developed for

Fig 1. Graphical illustration of our experimental pipeline for the DR-GPT large language model.

various classification tasks based on medical text, such as automatic extraction of significant findings in chest radiograph reports [1], disease ICD-code and treating department prediction based on patients' self-reports [2], and COVID-19 diagnosis based on chemosensory reports [3]. As such, the LLM-based approach shows promise for automatic determination of the DR classification label from unstructured text reports.

In the present study, we develop and evaluate a Finnish GPT-based system (DR-GPT) for auto-labeling the severity and gradability of diabetic retinopathy from clinician's reports with unstructured text format. In addition, we evaluate quantitatively the classification performance of the system using established measures and the utility of such system to generate weakly annotated data for training a convolutional neural network. As a result, we demonstrate that the DR-GPT can accurately determine the DR classification label from unstructured medical reports, and furthermore, that it can be used to annotate unlabeled sets of data to improve image-based classifiers with statistically significant difference compared to a standard supervised approach. A graphical illustration of our experimental pipeline is presented in Fig. 1.

Materials and methods

In this section we present the dataset used in our analyses, the data preprocessing methods, our experimental setup, and the measures used to evaluate the results.

Patient data

The research was conducted as a retrospective and registry-based study using a pseudonymized dataset of diabetic patient screening and follow-up studies as well as special healthcare visits at the Helsinki University Hospital (HUS) region over a period from 2016 to 2019. The dataset consists in total of 40236 studies from 31292 patients. During each visitation, both retinal fundi of a patient were photographed, and a physician examined the fundus images to describe the visible signs and severity of the patient's DR in unstructured medical reports using the Early Treatment Diabetic Retinopathy Study (ETDRS) grading system [4]. According to the Finnish law (Medical Research Act (488/1999) and Act on Secondary Use of Health and Social Data (552/2019)) and European General Data Protection Regulation (GDPR) rules 216/679, a retrospective and registry-based study does not need ethical permission or informed consent from subjects. The research permit was granted by the Helsinki University Hospital Chief Medical Officer (decision number 67/2020), Helsinki, Finland, July 1, 2020. The data was accessed March 4, 2021 and the authors did not have access to identifying information.

For the present study, one optician, one optometrist, and two specialists in ophthalmology have manually analyzed 26626 ($\approx 66.2\%$) of the reports, such that the ETDRS severity of both the left and right eye mentioned in each report was converted to the corresponding numeric level of the system. The reports that mentioned an ETDRS severity of at most 35 were annotated by the optician or the optometrist. The cases with a higher ETDRS severity and those cases with any abnormalities, such as other pathologies or ungradable reports, were evaluated by one of the two specialists in ophthalmology. The abnormal reports were also evaluated for ambiguity of the grade, i.e., if there was ambiguity regarding the severity of DR or which eye the DR severity corresponded to, and if visible laser scars or laser treatment was mentioned. The lateral

ambiguity prevents the determination of the correct label for each of the eyes, and the DR that manifests after laser treatment is not captured by the ETDRS scale because it alters the natural progression of diabetic retinopathy. The ambiguous and laser treated cases were assigned an auxiliary *Ungradable* label, while the fully gradable cases were assigned the *Gradable* label.

The dataset of patients were divided to model training, validation, and test sets with a greedy search algorithm that performed the division such that a) the reports from each patient could only reside in one of the sets, b) the distribution of ETDRS classes was to be as similar as possible in all of the sets, and c) the proportion of data in the sets was to be matched with 70%, 10%, and 20% for training, validation, and test sets, respectively. Avoiding the overlap of patients between the sets avoids overoptimistic results by the deep learning neural network memorizing possible spurious patient level patterns. The ETDRS distributions are presented in detail in the table in S1 Table.

We used a truncated version of the ETDRS (ETDRS-T), which includes the four least severe ETDRS classes, i.e., 10, 20, 35, and 43, and a single class that is a combination of the most severe ETDRS classes in our dataset, i.e., 47, 53A-D, 53E, 61, 65, and 71. This simplification of the grading scheme was applied to balance between the number of samples present in each class and the clinical relevance for separating class severity. Although ETDRS-T cannot differentiate the ETDRS classes ≥ 47 from one another, the cases with the ETDRS class 47 are among the first that can require treatment in the coming years, thus making this system suitable to be applied in healthcare. The class distributions of ETDRS-T in patient, report, and eye -level are presented in Table 1. In addition, we have also evaluated the DR-GPT on a binary DR classification system (RDR) used in previous studies [5, 6, 7]. The RDR system is defined as moderate or worse diabetic retinopathy on the proposed international diabetic retinopathy classification (ICDR) system [8], with ICDR classes lower than moderate DR assigned the label 0 and moderate or worse the label 1.

Table 1. ETDRS-T distributions for the medical report data.

Label	Development Set*			Test Set		
ETDRS-T / Gradable	Patients**	Reports†	Eyes	Patients**	Reports†	Eyes
10	14124	15563	29994	3506	3885	7507
20	1658	1934	2558	411	477	644
35	823	1093	1499	206	265	371
43	303	475	706	69	118	180
≥47	127	164	267	27	41	66
Gradable	15325	17512	35024	4219	4786	8786
Ungradable	2998	3487	6473	802	841	1694

* Development set denotes training and validation data.
** A patient is counted for a label if the label is mentioned in any of the reports of the patient.
† A report is counted for a label if the label is mentioned for either of the eyes.

We have observed that there were a number of duplicate reports in the dataset, despite them having been recorded in an unstructured manner. Specifically, there were a number of reports with identical contents for some of the patients with milder severity classes, e.g., the reports merely indicating that there was no DR to be observed. There were in total of 3818 such duplicate reports in the test set, with 3117 reports having ETDRS-T class 10 for both eyes, 351 reports with ETDRS-T class 20 for both eyes, 300 reports with ETDRS-T class 35 for both eyes, and 50 reports where there the severity was different for each eye, but at most ETDRS-T class 35. To examine the performance

of DR-GPT more in depth, we formed two test datasets. One of the sets was the test data as-is, i.e., following the empirical distribution of the reports, and the other set was formed with report stratification, i.e., keeping only one example of each unique report. The DR-GPT performance on the former set measures how well it performs on the population level, whereas the latter set measures how well the DR-GPT performs on a report level.

The medical reports are based on the fundus images taken during the patient visits. In a standard visitation, four 50° field-of-view fundus photographs are taken from both eyes. The four images consist of a macula centered color image and three red-free filtered images with one of them centered at the macula, one centered temporal to the macula, and one is centered nasal to the optic disc. However, some of the patient visits include more than four images per eye due to various reasons, e.g., patient requiring additional images due to previous laser treatment or patient requiring anterior segment images to visualise optic media opacities in detail. Additionally, some of the patients have missing images, e.g., when an eye cannot be imaged due to advanced eye disease or due to technical issues during photography. For our image classification experiments we included the eye image sets with exactly four images of a standard visitation to simplify the analysis. In total, the image-based experiments included annotations for 41738 eyes of 19293 patients from 22056 visitations and 18032 eyes of 7765 patients from 10026 visitations with no annotations. We used the same training, validation, and test splits for the patients as with the report classification experiments. Full description of the ETDRS-T class distributions for the image classification experiments is shown in Table 2.

Table 2. ETDRS-T distributions for the labelled images.

Label	Development Set*		Test Set	
ETDRS-T	Patients**	Eyes	Patients**	Eyes
10	14120	28594	3553	7210
20	1626	2441	405	622
35	790	1406	201	348
43	293	651	64	160
≥47	129	240	28	66

* Development set denotes training and validation data.
** A patient is counted for a label if the label is mentioned in any of the reports of the patient.

Data preprocessing

In order to limit the number of tokens, i.e., the text representations used as input to large language models, that require processing and to prevent the DR-GPT from learning spurious correlations between uninformative tokens and the class, we preprocessed the text data with multiple rules. The portions of the medical reports that consider observations regarding the patient's fundus are in a free-form text, whereas there are some automatically added parts that are structured. For example, the beginning and the end of the report have structured information, such as time and date of the clinical examination and the name of the examiner. Since this information does not consider the patient's health, we removed these parts from the texts. In addition, unnecessary characters added for visual purposes were automatically removed, e.g., multiple line-breaks. Finally, we utilized the text-tokenizer of TurkuNLP gpt3-finnish-small to convert the text strings to token indices for DR-GPT.

For our experiments with weakly annotated image data, we performed various preprocessing and data augmentation methods. The original fundus images were rectangular in shape with the fundus being visible as a circular region in the middle of the image and surrounded by black borders. We cropped each image to the smallest square image that contained the fundus entirely to remove as much of the black borders as possible. We then resized each image to a standard resolution of 512×512. During the training of the convolutional neural networks, we utilized training augmentations based on recent literature [5, 7, 9], i.e., random spatial flips both vertically and horizontally (p=0.5), random rotations uniformly within the range of $[-180°, 180°]$, random translations within the range of [-25,25] pixels in both spatial axes, and random zooms within range [90%,110%]. Finally, the image pixel values were mapped to the range [-1,1], during both the training and inference.

Deep learning models

As our large language model, we utilize the recently proposed Finnish GPT-3 model [10], namely the pretrained "Small" version of the architecture. The model is based on the BLOOM architecture [11], which is similar to the GPT-3, and the pretrained model has been trained in the next token prediction paradigm with various Finnish text resources. The model takes a sequence of tokens as an input and it outputs a sequence of probability distributions with the distribution at an index i representing the conditional probability of the token at index $i + 1$ given all the previous tokens.

The DR-GPT model consists of two Finnish GPT-3 neural networks, one of which classifies the degree of diabetic retinopathy on the ETDRS-T scale for the left and right eyes, and the other determines the gradability of the left and right eyes. In order to adapt the model for these classification tasks, we replace the final, i.e., the next token prediction, layer of the model with two parallel layers that predict the ETDRS-T level or gradability for each eye. Additionally, when feeding the tokenized text to the model, we only utilize the predictions on the last index of the sequence due to the causal masking used within the model, which ensures that all the text data is visible for the prediction. We use the cross-entropy loss function and regularize the models with 0.2 dropout rate that turned out to have the best performance from a grid-search with the values $[0.0, 0.1, \ldots, 0.9]$. All model parameters were fine-tuned with Adam optimizer [12] with learning rate 3×10^{-6} that was found to be the best in the range $[1 \times 10^{-3}, 1 \times 10^{-6}]$. We also examined AdamW optimizer [13] during our hyperparameter search, but as it turned out to perform similarly to Adam, we selected the latter for our main results.

As for the image classification experiment with weakly annotated data, we utilize an ImageNet [14] pretrained EfficientNet-B6 [15] convolutional neural network (CNN). In order to enable multi-view classification based on the four retinal images, we consider a combination of the sum and maximum multi-view fusion methods, described in detail in [16]. It first feeds each of the four fundus images to the CNN to create four output vectors. The output vectors are then processed by calculating the sum and maximum features in an element-wise manner across the four outputs. Lastly, the sum and maximum vectors are concatenated and fed to a multilayer perceptron with a softmax activation to output the conditional class probabilities. We slightly modified the original sum multi-view fusion approach by instead calculating the element-wise mean, as it turned out to be more numerically stable. We use the cross-entropy loss function, and AdamW optimizer with learning rate 3×10^{-5}, which we found to result in the best performance in the task.

Experiments and evaluation measures

To evaluate the classification performance of the five class ETDRS-T quantitatively, we use the quadratic weighted Cohen's kappa (QWK) [17], accuracy, and balanced accuracy measures. The QWK measure has previously been used in the evaluation of the deep learning algorithms for diabetic retinopathy classification with the five-class ICDR system [7, 18, 19], whereas the accuracy and balanced accuracy are common classification evaluation measures. For the binary classification tasks of RDR and gradability classification, we use the area under the receiver operating characteristic curve (AUROC), accuracy, and balanced accuracy. We trained 10 DR-GPTs and report the mean and standard deviation of the results with both empirical and report stratified distributions of test data.

The DR-GPT model outputs two vectors of probabilities that represent the conditional distribution of the classes for the left and right eyes given the text input. To calculate QWK, accuracy, and balanced accuracy, we select the label with the maximum probability as the prediction for each eye, i.e., $\hat{y}_{left} = \arg\max_c p(y_{left} = c \mid \boldsymbol{x})$ and $\hat{y}_{right} = \arg\max_c p(y_{right} = c \mid \boldsymbol{x})$. For multi-class classification with K categories, a set of N ground truth labels $\{y_1, \ldots, y_N\}$, and DR-GPT predicted labels $\{\hat{y}_1, \ldots, \hat{y}_N\}$, where the laterality of the label and prediction have been omitted for clarity, the QWK is defined as follows:

$$\text{QWK} = 1 - \frac{\sum_{i=1}^{K}\sum_{j=1}^{K}(i-j)^2 C_{i,j}}{\sum_{i=1}^{K}\sum_{j=1}^{K}(i-j)^2 E_{i,j}},$$

$$C_{i,j} = \sum_{n=1}^{N} I[y_n = i] \cdot I[\hat{y}_n = j],$$

$$E_{i,j} = \frac{1}{N}\sum_{a=1}^{K} C_{i,a} \sum_{b=1}^{K} C_{b,j},$$

where $I[\cdot]$ is the indicator function, C is the confusion matrix, and E the expected agreement matrix. Accuracy and balanced accuracy are defined as:

$$\text{Accuracy} = \frac{1}{N}\sum_{n=1}^{N} I[y_n = \hat{y}_n],$$

$$\text{Balanced Accuracy} = \frac{1}{K}\sum_{k=1}^{K} \frac{\sum_{n=1}^{N} I[y_n = k] \cdot I[\hat{y}_n = k]}{\sum_{n=1}^{N} I[y_n = k]}.$$

To obtain the binary classification measure AUROC, one needs to evaluate the sensitivity and specificity of a classifier at multiple probability thresholds τ, such that at each threshold the predicted label is defined as:

$$\hat{y} = \begin{cases} 1 & \text{if } p(y \mid \boldsymbol{x}) > \tau, \\ 0 & \text{else.} \end{cases}$$

The area under the curve defined by the sensitivities and specificities on thresholds $\tau \in [0, 1]$ is the AUROC measure. When evaluating the AUROC of RDR classification performance (RDR AUROC), we formed the probability of RDR positive by adding the probabilities of ETDRS-T classes 35, 43, and ≥ 47 together.

To evaluate if there are differences between an EfficientNet-B6 model trained with manually annotated data and one trained with the data augmented with DR-GPT weak annotations, we used the 10 DR-GPTs of medical report classification experiments to

generate weak annotations on the unlabelled set of 10026 reports. Specifically, we took an ensemble of the models and filtered out the reports that the ensemble DR-GPT predicted as being ungradable. For the rest of the reports, the maximum probability label was assigned as the target label. We trained 10 EfficientNet-B6 models on both manually annotated data and on manually annotated data augmented with weak annotations, and determined statistically significant differences with two-tailed Wilcoxon signed rank tests [20], by considering the p-values less than 0.05 as significant. To control the false discovery rate of multiple hypotheses, Benjamini-Hochberg procedure [21] was used to declare significant results, which accounted for the various classification measures used to compare the approaches. These classification measures were implemented in Python (3.9.12) using the scikit-learn (1.3.0) [22] and Wilcoxon signed rank tests were calculated using SciPy (1.8.0) [23].

Results

This section presents the results for the DR-GPT on classifying patients' medical reports, and for the EfficientNet-B6 on image classification with and without the DR-GPT generated weakly supervised data.

Classification of patient's medical reports

The results of ETDRS-T classification with the DR-GPT are presented in Table 3. On the empirical dataset, the DR-GPT reached 0.975 QWK, 0.987 accuracy, and 0.937 balanced accuracy in the ETDRS-T classification task, and 0.999 AUROC in the RDR classification. It turned out that the performance degraded slightly when evaluating the performance on the report stratified i.e., only a unique sentence, test data, the QWK value decreased to 0.962, accuracy to 0.952, and the balanced accuracy to 0.930. In addition, the RDR AUROC turned out to decrease to 0.994.

Table 3. ETDRS-T and RDR classification results of DR-GPT.

Test Set	QWK	Accuracy	Balanced Accuracy	RDR AUROC
Empirical	0.975 (0.002)	0.987 (0.001)	0.937 (0.006)	0.999 (0.000)
Stratified	0.962 (0.004)	0.952 (0.004)	0.930 (0.006)	0.994 (0.001)

In the case of binary classification for gradability, the DR-GPT resulted in an AUROC value of 0.996, accuracy value of 0.986, and balanced accuracy value of 0.971, when evaluating the full test data. The performance on the report stratified set of patient data showed that the performance decreases slightly with AUROC being 0.960, accuracy 0.952, and balanced accuracy 0.894. The results are illustrated in full on Table 4.

Table 4. Gradability classification results of DR-GPT.

Test Set	AUROC	Accuracy	Balanced Accuracy
Empirical	0.996 (0.000)	0.986 (0.000)	0.971 (0.003)
Stratified	0.960 (0.003)	0.952 (0.001)	0.894 (0.013)

In the case of the ensemble DR-GPT model, created by taking an ensemble of the 10 repetitions, the evaluation yielded 0.977 QWK, 0.988 accuracy, and 0.943 balanced accuracy in the ETDRS-T classification, and 0.999 AUROC in the RDR classification on the empirical set of test data. As for the report stratified test data, the DR classification performance resulted in 0.966 QWK, 0.957 accuracy, and 0.936 balanced

Fig 2. Ensemble DR-GPT confusion matrices on the empirical set of test data. **(a)** ETDRS-T classification, **(b)** RDR classification derived from the ETDRS-T, and **(c)** gradability classification.

accuracy in the ETDRS-T classification, and 0.995 AUROC in the RDR classification. In turn the ensemble DR-GPT model had gradability classification performance of 0.997 AUROC, 0.988 accuracy, and 0.974 balanced accuracy on the empirical test set, and 0.965 AUROC, 0.956 accuracy, and 0.903 balanced accuracy on the report stratified test set. The ensemble DR-GPT ETDRS-T and gradability confusion matrices, evaluated on the empirical set, are presented in Fig. 2. The DR-GPT ensemble predictions on the unlabelled data, used in the image classification experiments, resulted in 5627 patients with 9949 eyes in the ETDTS-T class 10, 1650 patients with 2429 eyes in class 20, 1460 patients with 2808 eyes in class 35, 922 patients with 2182 eyes in class 43, and 341 patients with 664 eyes in class ≥ 47.

Image classification with weak supervision

As for the image classification results with the EfficientNet-B6 CNN model, when trained with manually annotated data for ETDRS-T classification, it had the mean (standard deviation) value of 0.890 (0.005) for QWK, 0.924 (0.002) for accuracy, and 0.579 (0.030) for the balanced accuracy. In addition, when the model was evaluated for RDR classification, it had an RDR AUROC of 0.979 (0.003). It turned out that when the model was trained with the DR-GPT generated weak annotations, in addition to the manual annotations, the ETDRS-T classification performance improved to 0.892 (0.004) for QWK, 0.924 (0.003) for accuracy, and 0.604 (0.017) for balanced accuracy, in terms of the mean (standard deviation) of these measures. Furthermore, the RDR AUROC improved to 0.984 (0.001). Additionally, all the differences between the baseline supervised and the DR-GPT augmented model were statistically significant. A summary of the results are presented in Table 5 and the confusion matrices of ensembled models from both the experiments in Fig. 3.

Table 5. Diabetic retinopathy classification from fundus images.

Experiment	QWK	p-value	Accuracy	p-value	Balanced Accuracy	p-value	RDR AUROC	p-value
Supervised	0.886 (0.005)	0.003	0.922 (0.002)	0.002	0.642 (0.030)	0.014	0.979 (0.003)	1.8×10^{-4}
+ DR-GPT	0.893 (0.004)*		0.926 (0.002)*		0.667 (0.014)*		0.984 (0.001)*	

* Statistically significant differences ($p < 0.05$).

Fig 3. ETDRS-T and RDR confusion matrices for image classification with ensembles. **(a)** ETDRS-T with standard supervised approach, **(b)** ETDRS-T with DR-GPT weakly supervised data, **(c)** RDR with standard supervised approach, **(d)** RDR with DR-GPT weakly supervised data.

Discussion

In this work, we have proposed a large language model DR-GPT for automatic classification of diabetic retinopathy and its gradability from unstructured medical patient reports. We demonstrated that DR-GPT has a high accuracy on both of the classification tasks, and furthermore, that it can be used to automatically annotate fundus images by analysing the corresponding unlabeled medical reports. This weakly annotated data could in turn be used to augment the training data for a convolutional neural network to improve its performance.

Overall the DR-GPT model had excellent accuracy with only a few errors. An observation from the analysis is that the classification performance of all grading scales degraded systematically when evaluated with the report stratified dataset in comparison to the empirical dataset. This difference can be attributed to the duplicate reports being more abundant during training as well as having simpler and homogeneous content e.g., simply stating that there is no diabetic retinopathy, which makes their classification easier. Thus, the remaining performance gains for the DR-GPT could be achieved by increasing the number of the more severe cases, i.e., ETDRS-T classes 43 and ≥ 47, in the training set, as they have more heterogeneous reports with unique terminology.

As for the image classification experiments, the EfficientNet-B6 convolutional neural network with access to additional weakly supervised data generated by DR-GPT had a significant improvement across all the measures, when compared to the supervised baseline. This demonstrates that the DR-GPT approach can be used with practically no additional cost, besides the computation, to improve the performance of an image-based classifier. However, the size of the improvement was small or moderate across the measures, even when the weakly supervised CNN had access to additional images from 19492 eyes. This outcome can be due to the saturation of classification performance with a CNN based approach, and we expect that the DR-GPT weakly supervised data could improve upon the supervised baseline more when the labelled dataset of fundus images is small.

There are some limitations to our study. Firstly, the DR-GPT was trained using exclusively Helsinki University Hospital region data, which may differ in terms of how structured the medical patient reports are in comparison to other national regions. This poses a challenge in the generalization of DR-GPT to reports from other regions, which can be analyzed further with multi-center data. Secondly, there are medical reports that refer to the previous examinations that the patient has undergone, which were not available to the DR-GPT during training or inference. This lack of information could be remedied by architectural and data-processing development, which remains for the future work. Thirdly, we utilized the truncated ETDRS-T system instead of the ETDRS scale due to the limited number of cases with severe or worse background DR and proliferative DR in our dataset, which limits the applicability of the approach to more fine-grained analysis of DR. Lastly, our CNN experiments with DR-GPT generated weakly supervised data utilised a relatively simple multi-view fusion mechanism, which limited our analysis to the standard cases with exactly four images.

This pivotal study has shown the efficacy of LLMs with unstructured medical data and the synergy in combination with CNNs for image -based analysis with diabetic retinopathy screening. With the promise of the high accuracy of the method, a more robust analysis with multi-center and finer diagnostic grading scale is left for future work.

Conclusion

A large language model can accurately analyze diabetics' unstructured medical reports to identify the severity and gradability of diabetic retinopathy. It can be used to automatically generate structured diabetic retinopathy classification for existing databases and to generate training data for other deep learning -based models.

Supporting information

S1 Table. ETDRS distributions for the medical report data. A patient is counted for a label if the label is mentioned in any of the reports of the patient, and a report is counted for a label if the label is mentioned for either of the eyes. Development set denotes training and validation data.

References

1. Bressem KK, Adams LC, Gaudin RA, Tröltzsch D, Hamm B, Makowski MR, et al. Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. Bioinformatics. 2020;36(21):5255–5261. doi:10.1093/bioinformatics/btaa668.

2. Wang J, Zhang G, Wang W, Zhang K, Sheng Y. Cloud-based intelligent self-diagnosis and department recommendation service using Chinese medical BERT. Journal of Cloud Computing. 2021;10(1):4. doi:10.1186/s13677-020-00218-2.

3. Li H, Gerkin RC, Bakke A, Norel R, Cecchi G, Laudamiel C, et al. Text-based predictions of COVID-19 diagnosis from self-reported chemosensory descriptions. Communications Medicine. 2023;3(1):104. doi:10.1038/s43856-023-00334-5.

4. Davis MD, Fisher MR, Gangnon RE, Barton F, Aiello LM, Chew EY, et al. Risk factors for high-risk proliferative diabetic retinopathy and severe visual loss: Early Treatment Diabetic Retinopathy Study Report# 18. Investigative ophthalmology & visual science. 1998;39(2):233–252.

5. Leibig C, Allken V, Ayhan MS, Berens P, Wahl S. Leveraging uncertainty information from deep neural networks for disease detection. Scientific Reports. 2017;7(1):17816. doi:10.1038/s41598-017-17876-z.

6. Band N, Rudner TG, Feng Q, Filos A, Nado Z, Dusenberry MW, et al. Benchmarking Bayesian Deep Learning on Diabetic Retinopathy Detection Tasks. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2); 2021.

7. Jaskari J, Sahlsten J, Damoulas T, Knoblauch J, Särkkä S, Kärkkäinen L, et al. Uncertainty-Aware Deep Learning Methods for Robust Diabetic Retinopathy Classification. IEEE Access. 2022;10:76669–76681. doi:10.1109/ACCESS.2022.3192024.

8. Wilkinson C, Ferris FL, Klein RE, Lee PP, Agardh CD, Davis M, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology. 2003;110(9):1677–1682.

9. Farquhar S, Osborne MA, Gal Y. Radial Bayesian Neural Networks: Beyond Discrete Support In Large-Scale Bayesian Deep Learning. In: Chiappa S, Calandra R, editors. Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics. vol. 108 of Proceedings of Machine Learning Research. PMLR; 2020. p. 1352–1362. Available from: https://proceedings.mlr.press/v108/farquhar20a.html.

10. Luukkonen R, Komulainen V, Luoma J, Eskelinen A, Kanerva J, Kupari HM, et al.. FinGPT: Large Generative Models for a Small Language; 2023.

11. Scao TL, Fan A, Akiki C, Pavlick E, Ilić S, Hesslow D, et al.. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model; 2023.

12. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. In: International Conference on Learning Representations (ICLR); 2015.

13. Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv preprint arXiv:171105101. 2017;.

14. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009. p. 248–255.

15. Tan M, Le Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: Chaudhuri K, Salakhutdinov R, editors. Proceedings of the 36th International Conference on Machine Learning. vol. 97 of Proceedings of Machine Learning Research. PMLR; 2019. p. 6105–6114. Available from: http://proceedings.mlr.press/v97/tan19a.html.

16. Yan X, Hu S, Mao Y, Ye Y, Yu H. Deep multi-view learning methods: A review. Neurocomputing. 2021;448:106–129. doi:https://doi.org/10.1016/j.neucom.2021.03.090.

17. Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin. 1968;70(4):213–220. doi:10.1037/h0026256.

18. Krause J, Gulshan V, Rahimy E, Karth P, Widner K, Corrado GS, et al. Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. Ophthalmology. 2018;125(8):1264–1272. doi:10.1016/j.ophtha.2018.01.034.

19. Sahlsten J, Jaskari J, Kivinen J, Turunen L, Jaanio E, Hietala K, et al. Deep Learning Fundus Image Analysis for Diabetic Retinopathy and Macular Edema Grading. Scientific Reports. 2019;9(1):10750. doi:10.1038/s41598-019-47181-w.

20. Wilcoxon F. In: Individual Comparisons by Ranking Methods. New York, NY: Springer New York; 1992. p. 196–202. Available from: https://doi.org/10.1007/978-1-4612-4380-9_16.

21. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological). 1995;57(1):289–300. doi:https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

22. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830.

23. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods. 2020;17:261–272. doi:10.1038/s41592-019-0686-2.

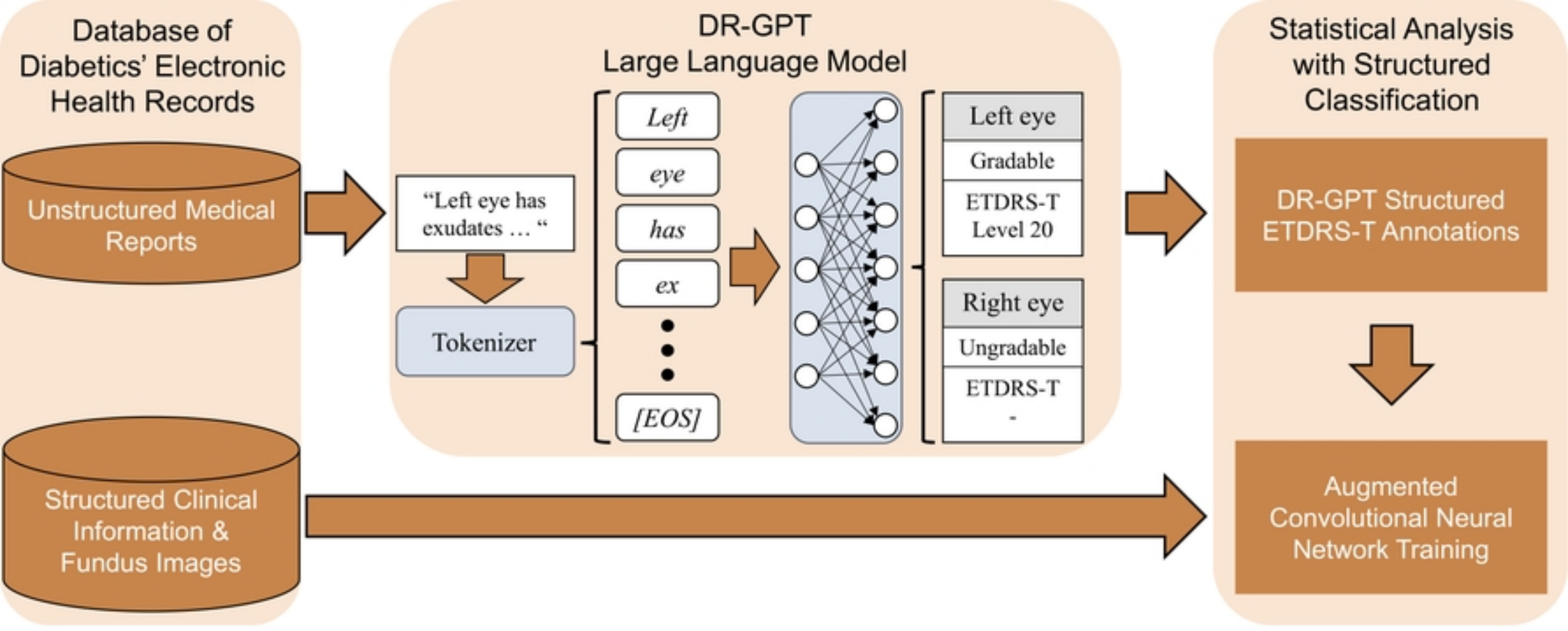
Database of Diabetics' Electronic Health Records
Unstructured Medical Reports
Structured Clinical Information & Fundus Images
DR-GPT Large Language Model
"Left eye has exudates … "
Tokenizer
Left
eye
has
ex
[EOS]
Left eye
Gradable
ETDRS-T Level 20
Right eye
Ungradable
ETDRS-T
-
Statistical Analysis with Structured Classification
DR-GPT Structured ETDRS-T Annotations
Augmented Convolutional Neural Network Training

Fig. 1

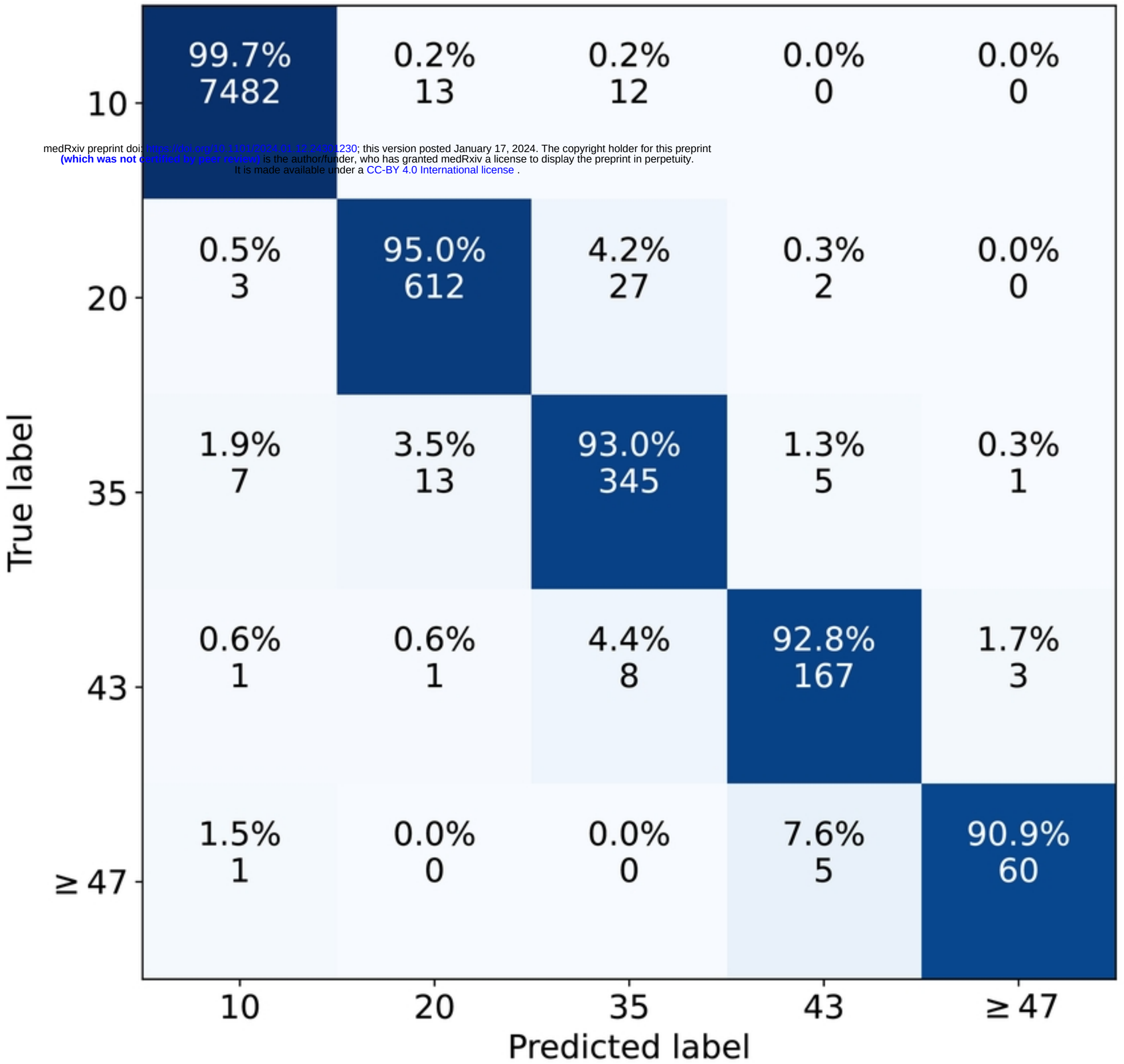

Fig. 2a

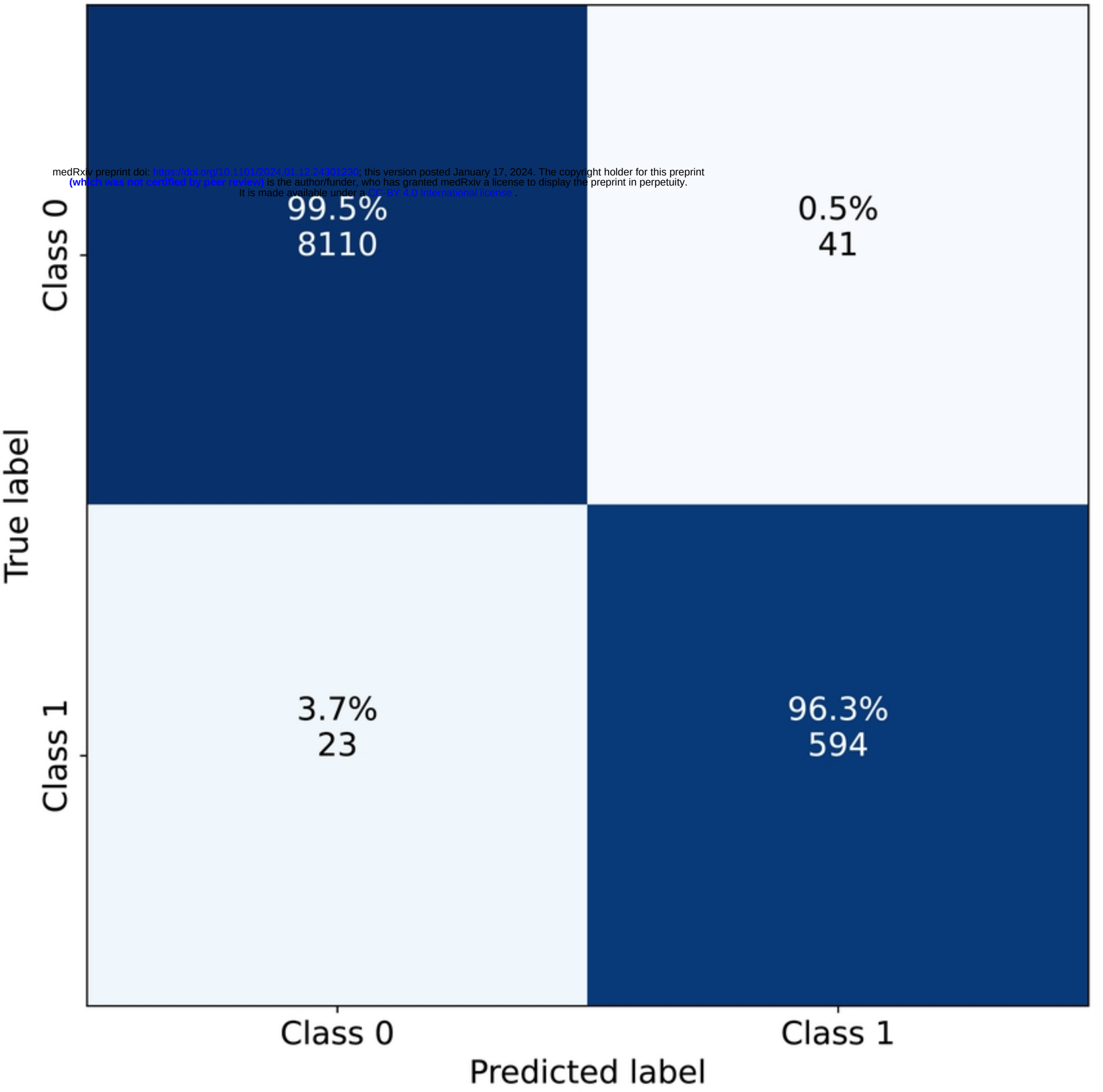

Fig. 2b

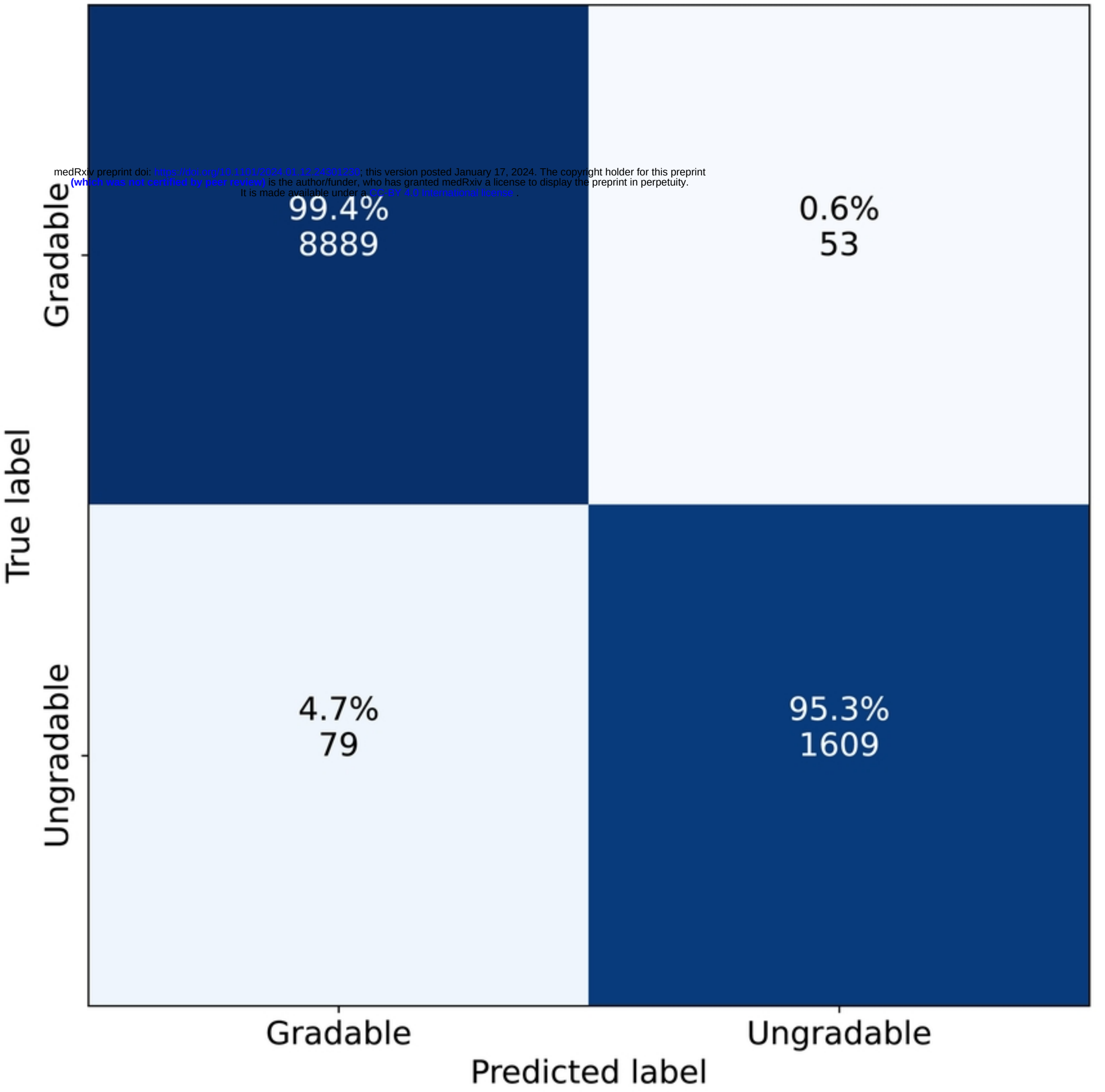

Fig. 2c

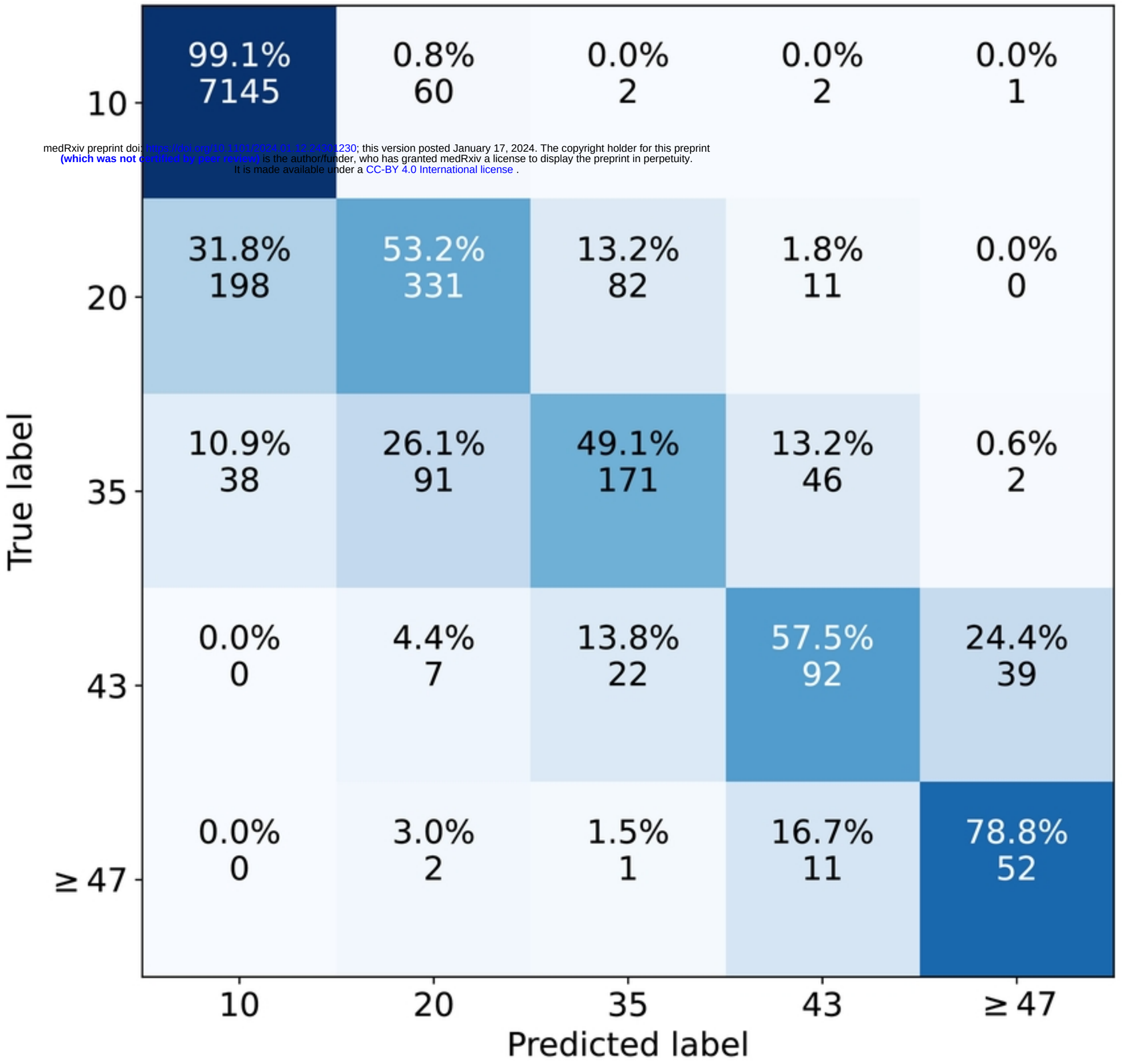

Fig. 3a

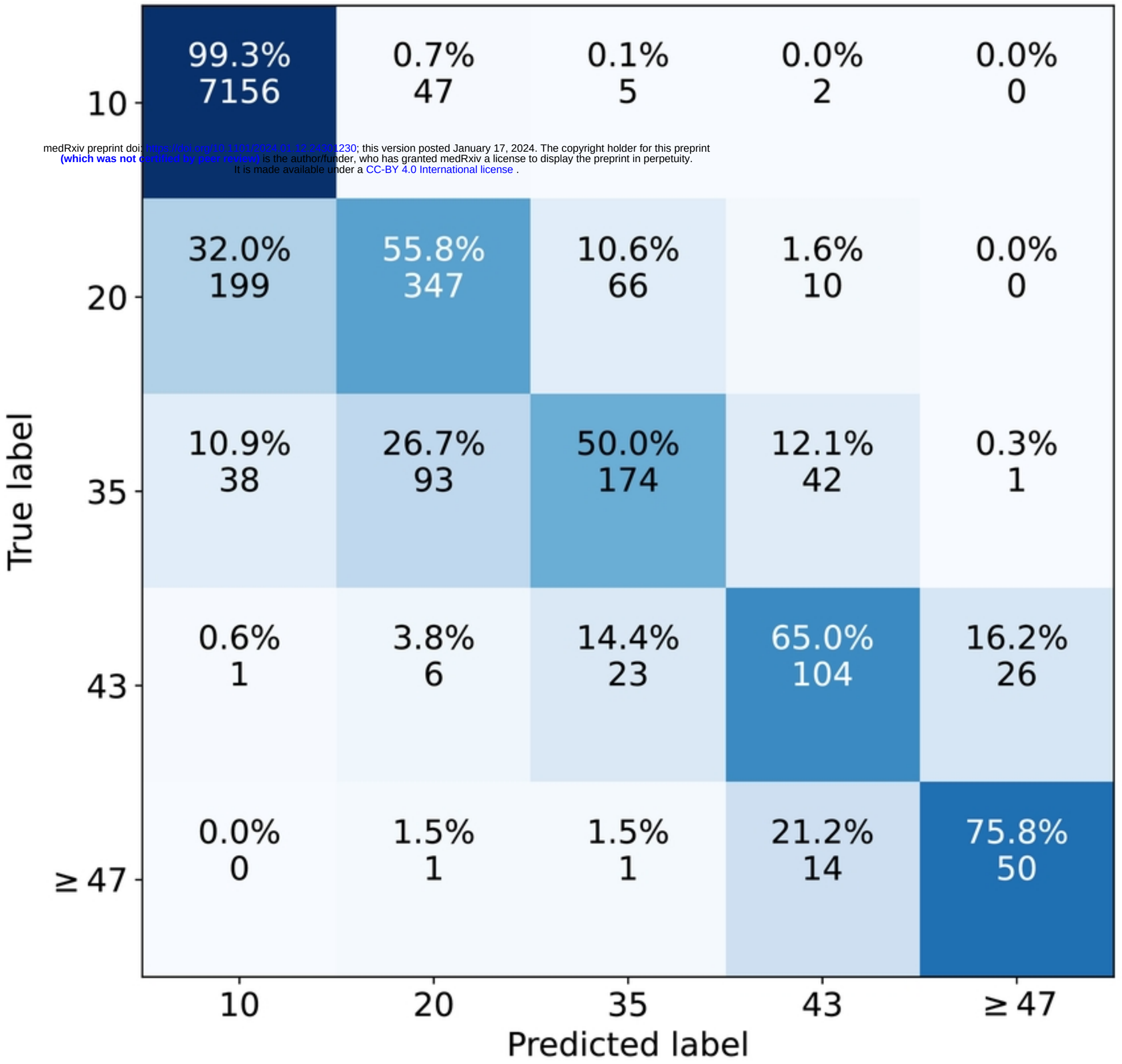

Fig. 3b

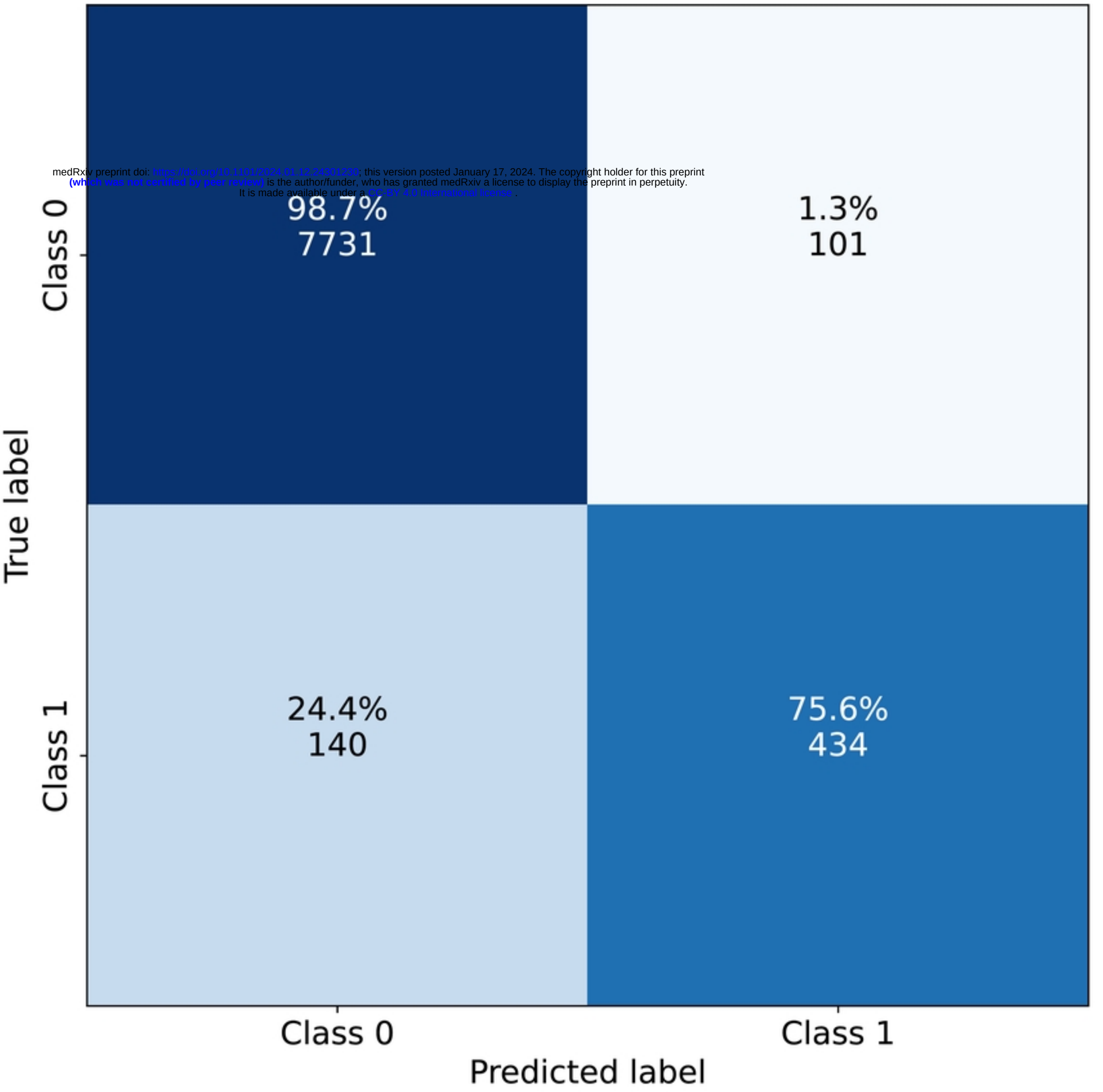

Fig. 3c

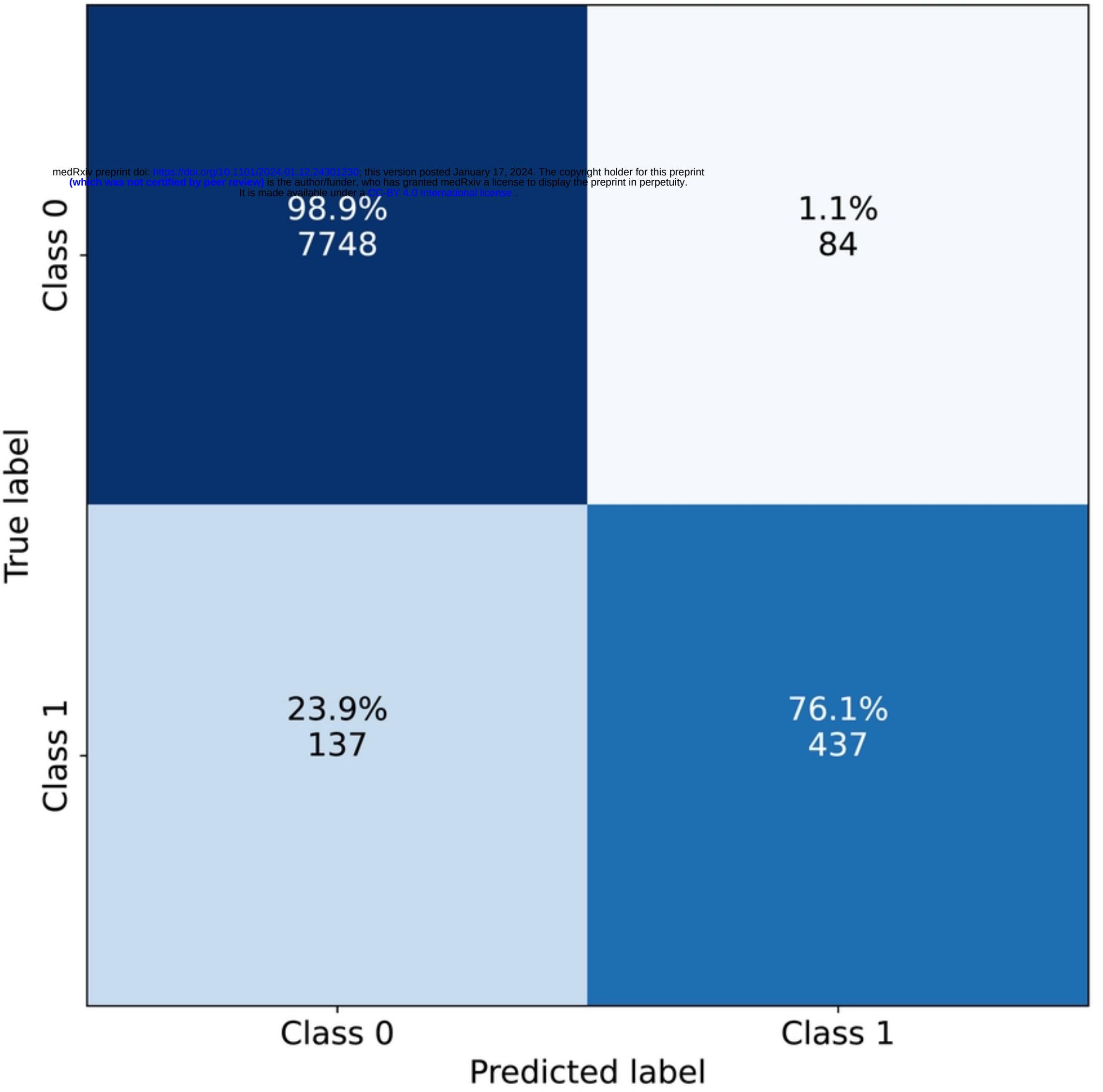

Fig. 3d