

Раздел III. Информатика

*Я.А. Пчелинцев¹, А.В. Хвостиков², О.П. Бучацкая³,
Н.А. Никифорова⁴, Л.П. Шепелева⁵, Е.С. Прокопьев⁶,
Л.Е. Паролина⁷, А.С. Крылов⁸*

АНАЛИЗ РОБАСТНОСТИ КОМПЬЮТЕРНОЙ ДИАГНОСТИКИ ТУБЕРКУЛЕЗА ПО РЕНТГЕНОВСКИМ СНИМКАМ ГРУДНОЙ КЛЕТКИ

Введение

Свёрточные нейронные сети произвели революцию в сфере обработки изображений, и одной из областей, на которой это отразилось, является медицина. Хотя алгоритмы глубокого обучения требуют большого количества данных для обучения, с чем могут возникать сложности из-за требований конфиденциальности и ограничений со стороны закона на использование полностью автоматизированных систем диагностики и обработки медицинских данных, совершается всё больше попыток внедрить автоматические алгоритмы в медицинскую практику. Одной из областей медицины, в значительной мере опирающихся на визуальные данные, является анализ рентгеновских снимков грудной клетки, что включает в себя диагностику туберкулёза [1, 2, 3, 4].

Однако, в отличие от таких заболеваний лёгких, как бактериальная или вирусная пневмония, данных по диагностике туберкулёза в открытом доступе не так много. Кроме того, из-за частичной несовместимости различных наборов рентгеновских изображений грудной клетки качество работы алгоритмов диагностики на рентгенограммах из наборов, не попавших в обучающую выборку, может сильно падать. Причины подобной низкой робастности (то есть низкой устойчивости качества работы алгоритма с различными наборами входных данных) включают в себя различия в условиях получения снимков, в диапазонах и методах преобразования интенсивности пикселей, в сложности визуального разделения классов. В частности, определение уровня жёсткости

¹ факультет ВМК МГУ имени М.В. Ломоносова, e-mail: yapchelintsev@live.cs.msu.ru.

² факультет ВМК МГУ имени М.В. Ломоносова, e-mail: khvostikov@cs.msu.ru.

³ факультет ВМК МГУ имени М.В. Ломоносова, e-mail: buchatskaya_olga_112@mail.ru.

⁴ ГБУ РС(Я) НПЦ "Фтизиатрия" имени Е.Н. Андреева, e-mail: nikiada@mail.ru.

⁵ ГБУ РС(Я) НПЦ "Фтизиатрия" имени Е.Н. Андреева, e-mail: shepelevalp@mail.ru.

⁶ ГБУ РС(Я) НПЦ "Фтизиатрия" имени Е.Н. Андреева, e-mail: Prokopeves@ftiz14.ru.

⁷ «НМИЦ фтизиопульмонологии и инфекционных заболеваний» Минздрава России, e-mail: parolina@nmrc.ru.

⁸ факультет ВМК МГУ имени М.В. Ломоносова, e-mail: kryl@cs.msu.ru.

рентгеновских снимков грудной клетки и его влияние на качество диагностики туберкулёза было рассмотрено в [5, 6], где было показана возможность повышения качества работы алгоритма классификации путём контролируемого формирования обучающей выборки. Анализ доменного сдвига (явления, при котором наблюдается снижение качества работы алгоритмов при его замерах на наборах данных, относящихся к той же области, но отличных от обучающего) для рентгенограмм органов грудной клетки был проведён в [7] между некоторыми открытыми наборами применительно к задаче обнаружения изображения лёгких и в [8] – для крупнейших открытых наборов (наборов, находящихся в открытом доступе) применительно к задаче диагностики патологии легких.

Общедоступные данные

Согласно недавним обзорам методов диагностики туберкулёза на основе глубокого обучения, таким как [1, 2, 3], основными открытыми источниками данных для задачи диагностики туберкулёза по рентгеновским изображениям грудной клетки всё ещё являются наборы Montgomery County [9] и Shenzhen [10]. Другие репрезентативные наборы снимков в свободном доступе [4] включают DA и DB [11] и TBX11K [12].

Набор Montgomery County (далее MC) содержит изображения в формате PNG с разрешением 4020x4892 пикселей и глубиной цвета 8 бит. Набор Shenzhen (далее SZ) содержит изображения в формате PNG с различным разрешением (примерно 3000x3000 пикселей) и глубиной цвета 8 бит. Так как эти наборы в большинстве работ используются совместно [1, 2], то было принято решение их объединить.

Набор DA содержит изображения в формате JPG с разрешением 1024x1024 или 2320x2828 пикселей и глубиной цвета 8 бит. Набор DB содержит изображения в формате DICOM с разрешением 3008x3008 пикселей и глубиной цвета 16 бит, и в рамках данного исследования был преобразован в формат PNG с глубиной цвета 8 бит. Эти наборы также были объединены в силу их родства и малого размера по отдельности. Набор DB был переведён в формат PNG с глубиной цвета 8 бит.

Набор TBX11K содержит изображения в формате PNG с разрешением 512x512 пикселей и глубиной цвета 8 бит.

Примеры изображений из этих наборов (обрезанные до равного соотношения сторон) приведены на Рис. 1-3.

Фактическое количество изображений лёгких здоровых пациентов (класс Healthy) и изображений лёгких пациентов с проявлениями туберкулёза (класс TB), для которых имеются метки классов, в этих наборах представлены в Табл. 1. Некоторые снимки больных туберкулёзом из набора DB отсутствуют в его репозитории на сайте SourceForge. Что касается TBX11K, то он изначально поделен на обучающую,

валидационную и тестовую выборку, при этом последняя не имеет разметки и потому не вошла в итоговое количество снимков в таблице.



а) здоровые пациенты (класс Healthy)



б) пациенты, больные туберкулёзом (класс TB)

Рис. 1. Примеры изображений из объединённого набора MC и SZ, соответствующие разным классам

Табл. 1. Размеры рассматриваемых открытых наборов рентгеновских изображений грудной клетки

Набор данных	Класс		Количество снимков
	TB	Healthy	
Montgomery County (MC)	58	80	138
Shenzhen (SZ)	336	326	662
DA	78	78	156
DB	47	75	122
TBX11K	800	3800	4600

Каждый из этих наборов довольно однороден по уровню жёсткости, типу оборудования и общему качеству изображений. Однако замеры качества работы алгоритмов классификации между этими наборами могут демонстрировать существенную разницу в точности. Поэтому, как будет показано далее, модель, обученная, например, на тренировочной выборке самого объемного набора TBX11K с такими мерами предотвращения переобучения, как аугментация обучающей выборки, постепенное уменьшение величины шага градиентного спуска и ранняя остановка, имея значения чувствительности и специфичности на валидационной и тестовой выборке этого набора выше 0.95, на объединённом наборе MC + SZ демонстрирует падение качества до уровня ниже того, который можно получить при обучении непосредственно на самом наборе MC + SZ. В частности, влияние различия в распределении пациентов по возрасту на

качество сегментации грудной клетки на рентгеновских снимках упоминается в [13], а влияние гендерного дисбаланса на точность классификации рассмотрено в [14].



а) здоровые пациенты (класс Healthy)



б) пациенты, больные туберкулёзом (класс TB)

Рис. 2. Примеры изображений из объединённого набора DA и DB, соответствующие разным классам



а) здоровые пациенты (класс Healthy)



б) пациенты, больные туберкулёзом (класс TB)

Рис. 3. Примеры изображений из набора TBX11K, соответствующие разным классам

Благодаря своему большому размеру набор TBX11K представляется хорошим кандидатом на роль основы для обучающей выборки, так как он позволяет нейронной сети сформировать большое количество свёрточных фильтров, специфичных для задач обработки рентгеновских снимков грудной клетки. Однако так как на нём слишком легко достичь высоких результатов качества диагностики (будет продемонстрировано ниже), то

его следует дополнить более сложными с точки зрения задачи диагностики данными, чтобы покрыть больше возможных случаев активного туберкулёза с малым и большим объёмом поражения лёгких. Такие дополняющие данные должны состоять из различных по визуальным свойствам изображений, в противном случае на практике применимость получающегося алгоритма будет всё ещё ограниченной. По этой причине представляется целесообразным составить набор рентгеновских изображений грудной клетки различного качества, с различными условиями получения данных, сложностью обнаружения туберкулёза, который представлял бы широкий спектр встречающихся на практике видов патологических изменений.

Сбор данных

Чтобы решить проблему нехватки рентгеновских изображений органов грудной клетки больных туберкулёзом пациентов, мы представляем набор снимков, сделанных во фронтальной проекции, собранных в результате сотрудничества с несколькими медицинскими учреждениями Республики Саха (Якутия). Процедура формирования набора изображений из имеющихся коллекций рентгенограмм включала в себя удаление снимков с боковой проекцией; контроль правильности установленных значений «Photometric Interpretation» в файлах DICOM; удаление повторяющихся сессий; удаление пациентов с несколькими сессиями и различиями в диагнозе между сессиями; удаление несовершеннолетних пациентов; а также некоторую балансировку соотношения полов и распределений возрастов и диагнозов. В результате было отобрано 400 изображений здоровых пациентов и 400 изображений пациентов, больных туберкулёзом, где каждому пациенту соответствует только 1 снимок. Для получения рентгенограмм использовались стационарные и переносные комплексы оборудования. В основном разрешение изображений примерно равно 3000x3000, но часть снимков имеет меньший размер вплоть до около 2000x2000. Большинство изображений имеет глубину цвета 16 бит, у остальных снимков она равна 8 битам.

Врачебная диагностика осуществлялась посредством независимого двойного чтения снимков с подтверждением диагноза «туберкулёз» специалистами НПЦ «Фтизиатрия» им. Е.Н. Андреева на основании клинико-лабораторных и микробиологических данных.

Примеры рентгенограмм из сформированного набора изображений представлены на Рис. 4. Сравнение распределений снимков по диагнозу, полу и возрасту полученного набора и набора MC + SZ (для остальных наборов такие данные недоступны) изображены на Рис. 5-7. Сам набор изображений доступен для загрузки на странице по ссылке: <https://imaging.cs.msu.ru/en/research/TB/Sakha-TB>.



а) здоровые пациенты (класс Healthy)



б) пациенты, больные туберкулёзом (класс TB)

Рис. 4. Примеры изображений из набора Sakha-TB, соответствующие разным классам

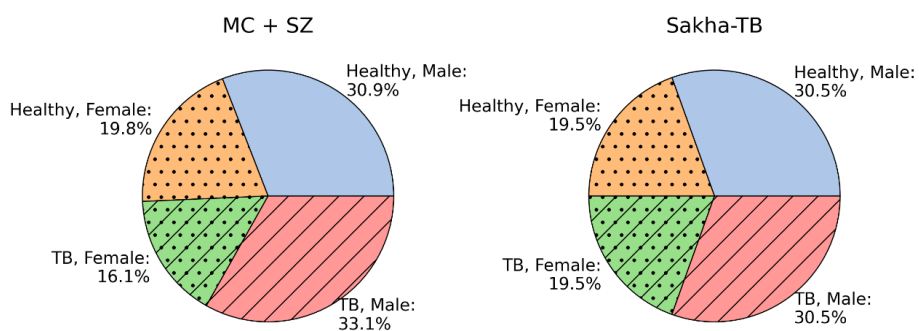


Рис. 5. Распределение изображений в наборах MC + SZ и Sakha-TB по диагнозу и полу

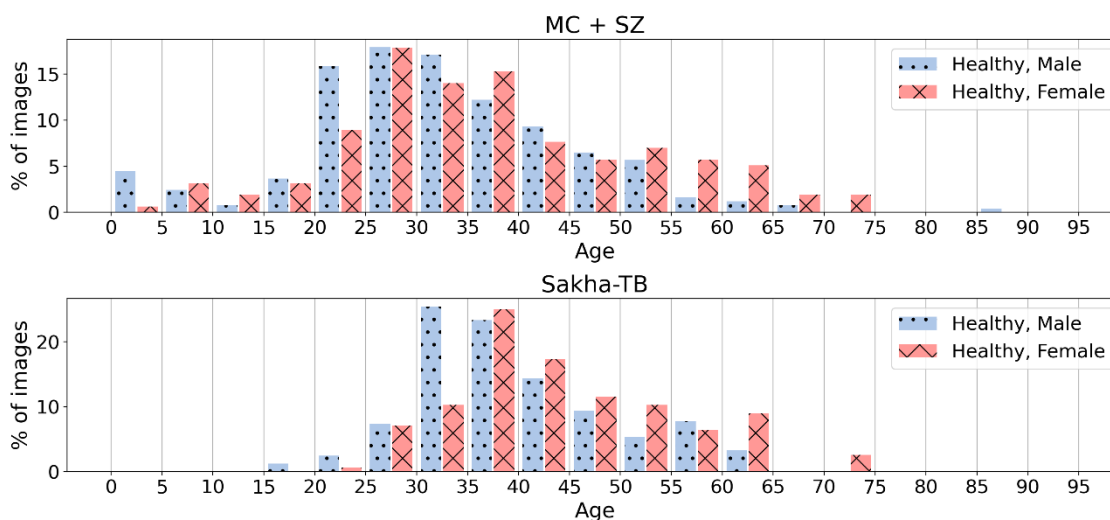


Рис. 6. Распределение изображений здоровых пациентов в наборах MC + SZ и Sakha-TB по возрасту в зависимости от пола

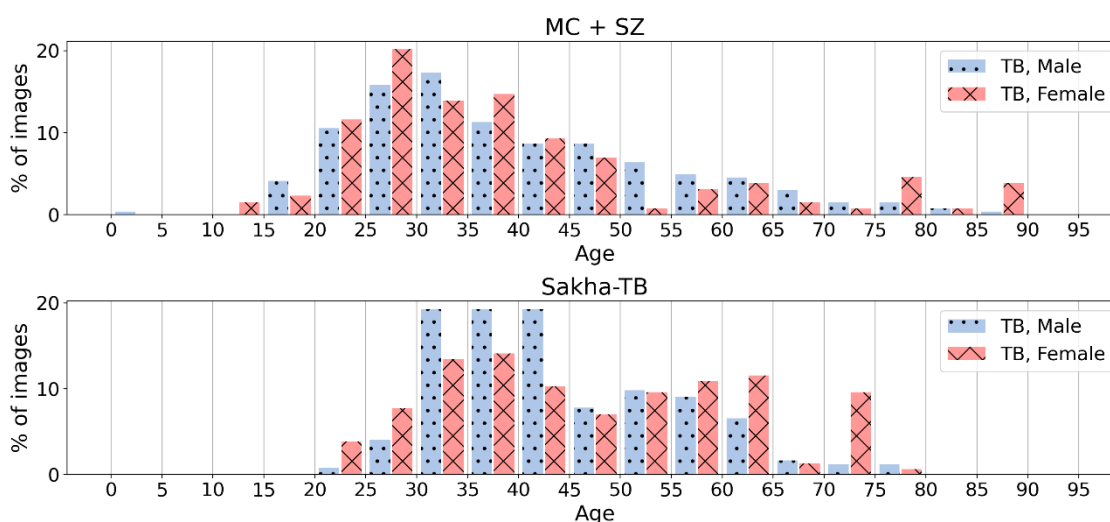


Рис. 7. Распределение изображений больных туберкулёзом пациентов в наборах MC + SZ и Sakha-TB по возрасту в зависимости от пола

Эксперименты

Для оценки возможностей применения предлагаемого набора изображений в задаче компьютерной диагностики туберкулёза были сделаны замеры качества работы методов глубокого обучения при обучении и тестировании на разных комбинациях рассматриваемых открытых наборов снимков:

1. MC + SZ,
2. DA + DB,
3. TBX11K,
4. Sakha-TB,
5. (MC + SZ) + Sakha-TB,

6. (DA + DB) + Sakha-TB,
7. TBX11K + Sakha-TB,
8. TBX11K + (MC + SZ),
9. TBX11K + (MC + SZ) + (DA + DB),
10. TBX11K + (MC + SZ) + (DA + DB) + Sakha-TB.

Глубина цвета снимков из набора Sakha-TB предварительно была приведена к 8 битам, а разрешение путём усреднения соседних пикселей было уменьшено в целое число раз до достижения значений около 1000x1000. Одновременно была проведена нормализация диапазона интенсивности пикселей с отсечением (англ. clipping) по 0.5% пикселей с каждого конца гистограммы изображения и обрезкой по 1.5% пикселей с каждой стороны изображений, чтобы увеличить долю диапазона допустимых значений 8 бит информации, приходящуюся на внутренние органы грудной клетки.

Для обучения были выбраны 2 современные нейросетевые модели, которые демонстрируют высокую точность в задачах классификации изображений: EfficientNetV2-M [15] и VisionTransformer-Base/16 [16], далее они будут обозначаться EffNetV2-M и ViT-B16 соответственно. Веса обеих сетей во всех случаях были инициализированы весами для решения задачи классификации изображений ImageNet [17], во время дальнейшего обучения были задействованы все слои.

Все базовые наборы изображений и их рассмотренные комбинации делились на обучающую, валидационную и тестовую выборки в соотношении 64:16:20 с предварительным случайным перемешиванием изображений и стратификацией по диагнозу для сохранения пропорций между классами. Изображения обучающей выборки в процессе обучения подвергались случайным преобразованиям, а именно:

- поворотам (в пределах 15 градусов в каждую сторону),
- масштабированию (коэффициент выбирался случайно из отрезка [0.8, 1.2]),
- сдвигам (до 30% размера изображения по каждой оси),
- изменению яркости и контраста (до 20% в каждую сторону).

В качестве функции потерь в процессе обучения и валидации использовалась кросс-энтропия:

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

а в качестве меры точности диагностики на валидационной выборке – сбалансированная точность [18].

Этап предобработки входного изображения заключался в последовательности следующих шагов:

1. автоматическое контрастирование изображения:

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

где $p_{0.5}(x)$ и $p_{99.5}(x)$ – 0.5%- и 99.5%-процентили значений интенсивности пикселей изображения;

2. автоматическое гамма-преобразование интенсивности пикселей изображения:

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

где x – значение интенсивности пикселя входного изображения, μ – средняя интенсивность всего изображения;

3. уменьшение изображения до входного разрешения, используемого нейронной сетью (512x512 пикселей для модели EffNetV2-M и 384x384 пикселей для ViT-B16).

Для минимизации функции потерь применялся алгоритм AdamW [19] с параметрами $lr = 5 \cdot 10^{-6}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\lambda = 0.01$. Размер пакета (англ. batch) изображений составлял 64 для модели EffNetV2-M и 32 для ViT-B16. В конце каждой эпохи качество модели замерялось на валидационной выборке, при отсутствии уменьшения значения функции потерь на ней в течение 5 эпох шаг градиентного спуска уменьшался в 5 раз. Для предотвращения переобучения при отсутствии уменьшения значения функции потерь на валидационной выборке в течение 16 эпох обучение прерывалось. В качестве финального состояния весов модели выбирались веса, соответствующие эпохе с наивысшим значением метрики качества на валидационной выборке. Графики зависимости функции потерь от номера эпохи в процессе обучения представлены на Рис. 8-11.

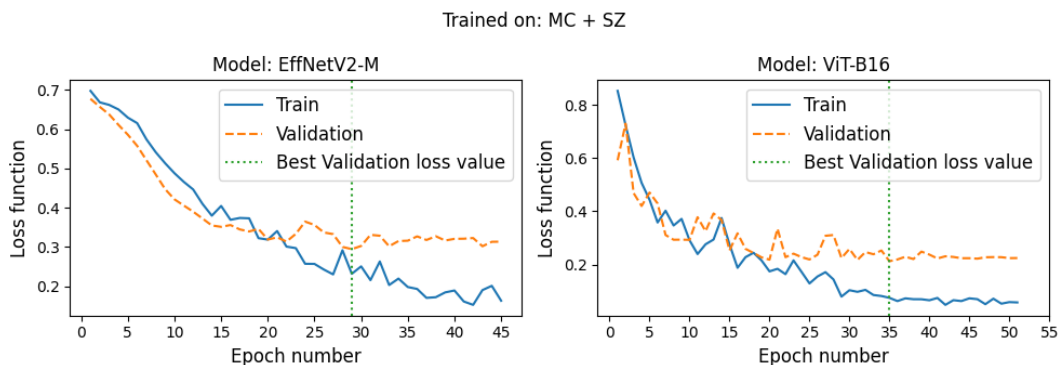


Рис. 8. Графики зависимости функции потерь от номера эпохи для набора данных MC + SZ

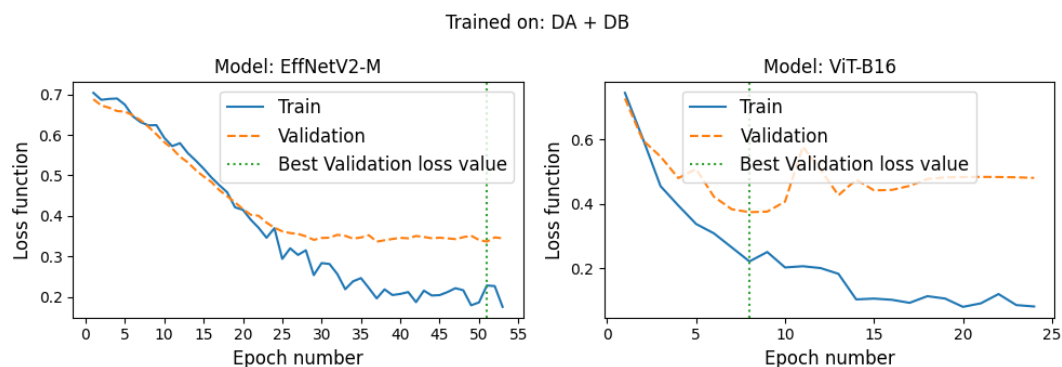


Рис. 9. Графики зависимости функции потерь от номера эпохи для набора данных DA + DB

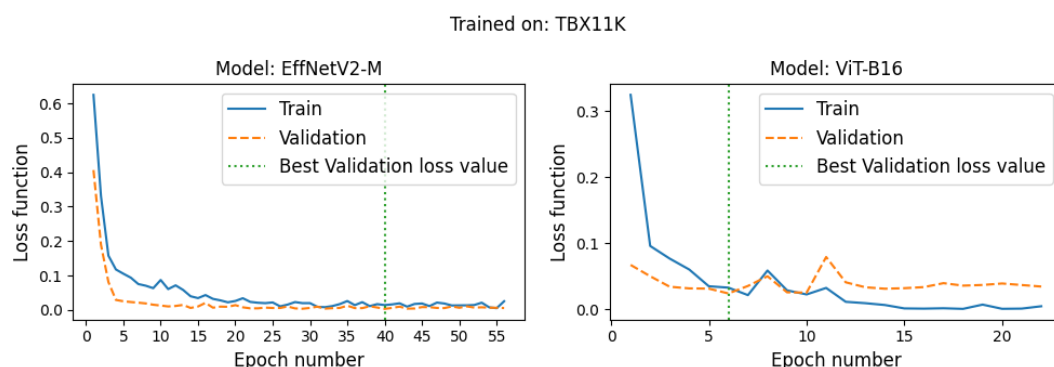


Рис. 10. Графики зависимости функции потерь от номера эпохи для набора данных TBX11K

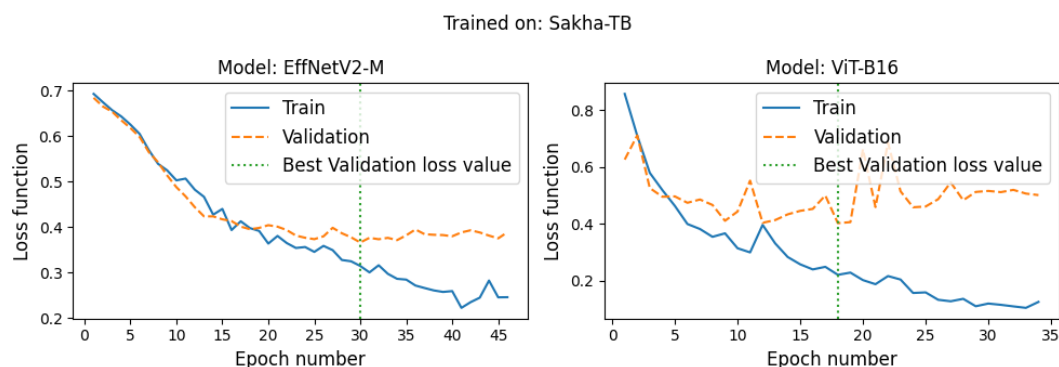
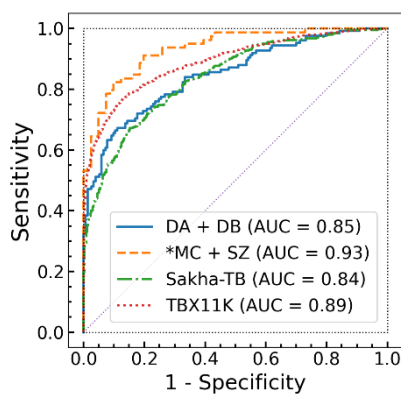


Рис. 11. Графики зависимости функции потерь от номера эпохи для набора данных Sakha-TB

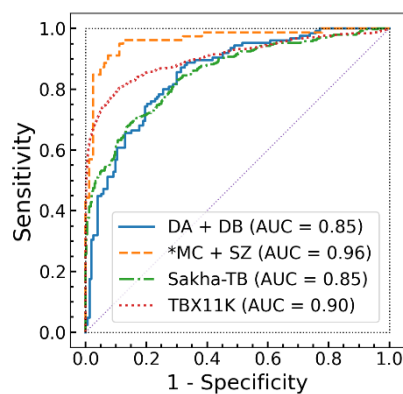
Результаты

Итоговые замеры качества работы обученных алгоритмов на рассматриваемых наборах производились с использованием широко используемых в медицине таких статистических показателей, как чувствительность (англ. *sensitivity*) и специфичность (англ. *specificity*). Были построены ROC-кривые, которые позволяют оценить общие

возможности алгоритма классификации с учётом возможности варьирования порога уверенности для каждого класса. Интегральным показателем возможностей алгоритма в таком случае может считаться величина ROC AUC, представляющая собой площадь под ROC-кривой рассматриваемого алгоритма, которая принимает значения от 0 до 1, где 1 соответствует абсолютно точному алгоритму. ROC-кривые с показателями ROC AUC для рассмотренных моделей, обученных на приведённых выше комбинациях рассматриваемых наборов рентгеновских снимков грудной клетки, представлены на Рис. 12-21. Для построения кривых для наборов, вошедших в обучающую выборку алгоритма, использовались их тестовые выборки, такие наборы выделены звёздочкой в описании графика; в противном случае использовались все наборы целиком.

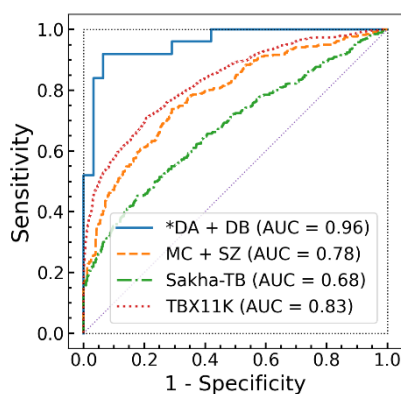


a) EffNetV2-M

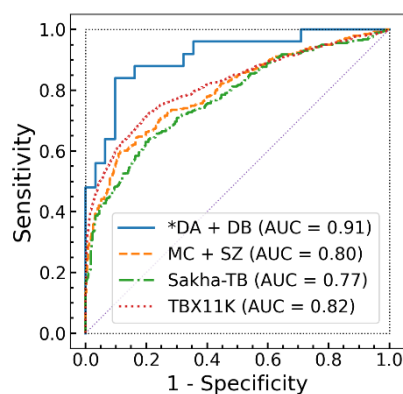


б) ViT-B16

Рис. 12. Графики ROC-кривых для моделей, обученных на наборе MC + SZ

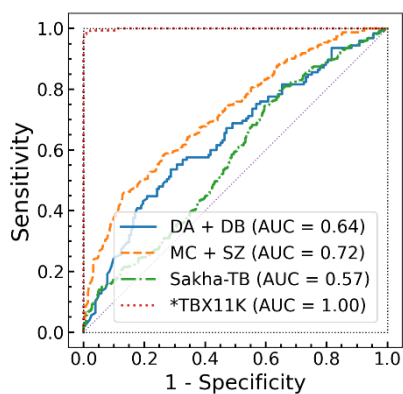


a) EffNetV2-M

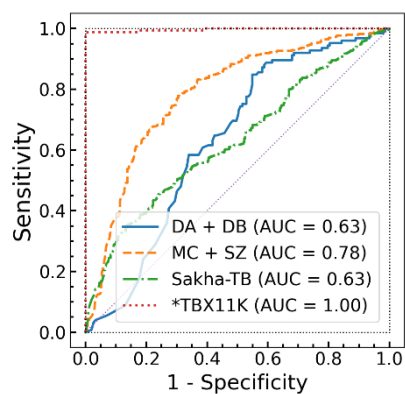


б) ViT-B16

Рис. 13. Графики ROC-кривых для моделей, обученных на наборе DA + DB

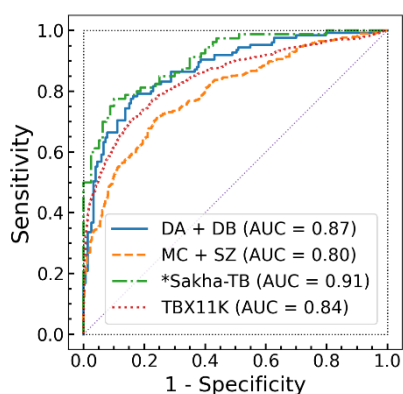


a) EffNetV2-M

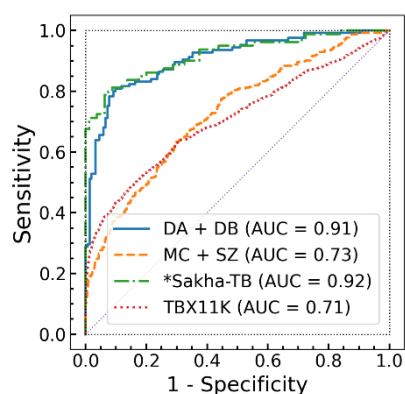


б) ViT-B16

Рис. 14. Графики ROC-кривых для моделей, обученных на наборе TBX11K

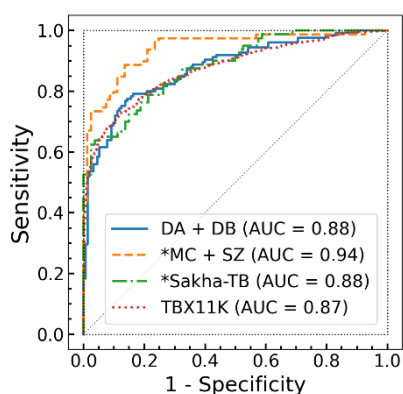


a) EffNetV2-M

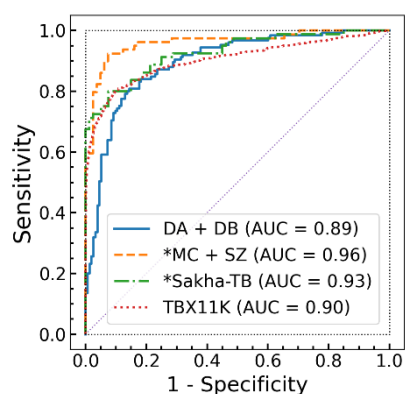


б) ViT-B16

Рис. 15. Графики ROC-кривых для моделей, обученных на наборе Sakha-TB

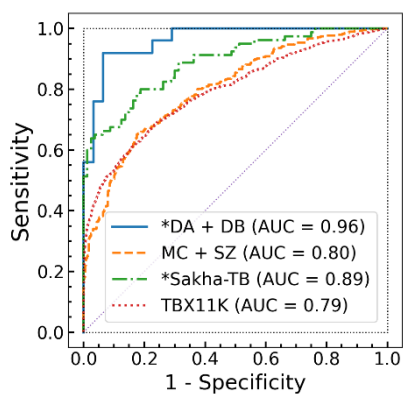


a) EffNetV2-M

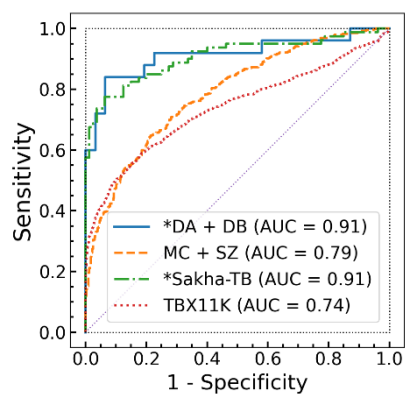


б) ViT-B16

Рис. 16. Графики ROC-кривых для моделей, обученных на наборе (MC + SZ) + Sakha-TB

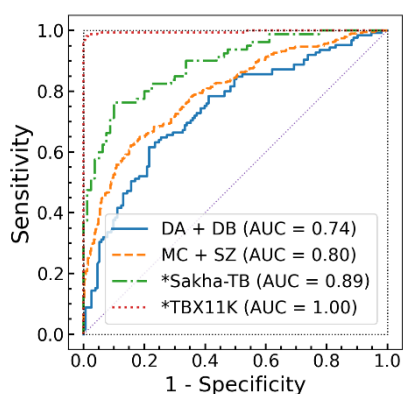


a) EffNetV2-M

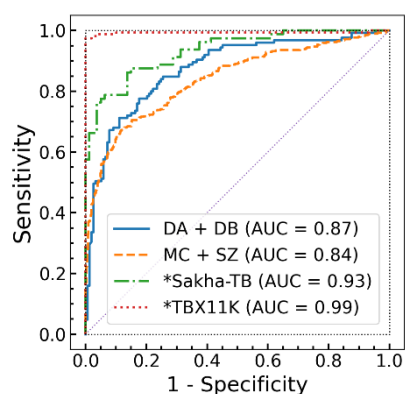


б) ViT-B16

Рис. 17. Графики ROC-кривых для моделей, обученных на наборе (DA + DB) + Sakha-TB

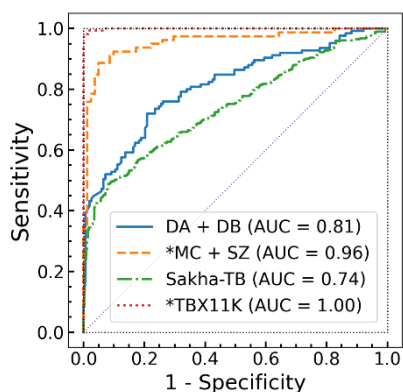


a) EffNetV2-M

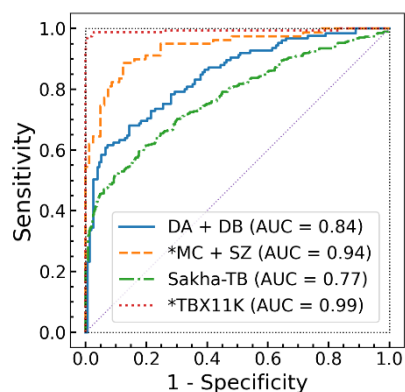


б) ViT-B16

Рис. 18. Графики ROC-кривых для моделей, обученных на наборе TBX11K + Sakha-TB

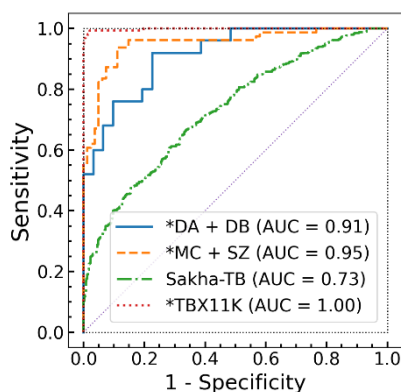


a) EffNetV2-M

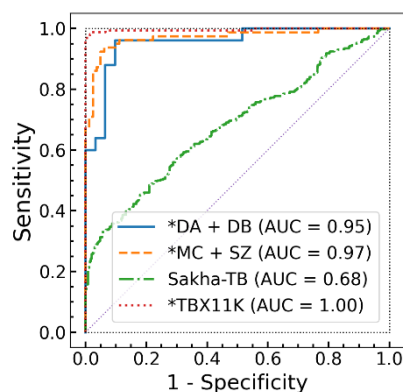


б) ViT-B16

Рис. 19. Графики ROC-кривых для моделей, обученных на наборе TBX11K + (MC + SZ)

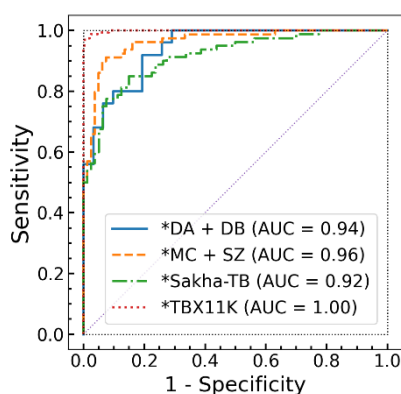


а) EffNetV2-M

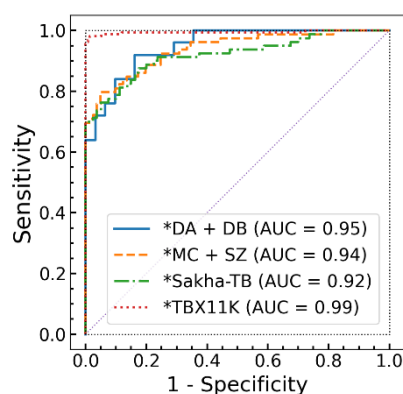


б) ViT-B16

Рис. 20. Графики ROC-кривых для моделей, обученных на наборе TBX11K + (MC + SZ) + (DA + DB)



а) EffNetV2-M



б) ViT-B16

Рис. 21. Графики ROC-кривых для моделей, обученных на наборе TBX11K + (MC + SZ) + (DA + DB) + Sakha-TB

Хотя результаты могут значительно различаться в зависимости от использованного алгоритма классификации, на основании графиков и значений меры ROC AUC на Рис. 12-15 можно сделать вывод, что среди базовых наборов изображений наиболее «представительным» являлся MC + SZ, так как позволил достичь более высокого качества работы на остальных наборах и в том числе на предлагаемом наборе (лучшее значение ROC AUC среди 2 моделей – не ниже 0.85). В то же время TBX11K слабо готовил алгоритм к работе на данных, отличных от подобных ему. Предлагаемый нами набор проявил себя лучше, чем DA + DB, но хуже, чем MC + SZ.

Комбинации 2 наборов изображений во время обучения позволил в некоторой степени закрыть «пробелы» каждого из наборов, как видно на Рис. 16-19. Лучше всего себя показала комбинация среднего по размеру MC + SZ с предлагаемым Sakha-TB, позволяя нейросетевому алгоритму

почти в равной степени хорошо работать на всех 4 базовых наборах; а сочетание TBX11K и MC + SZ всё так же плохо диагностировало заболевание на оставшихся 2 наборах, вероятно, из-за дисбаланса в размерах обеих компонент комбинации.

Объединение всех 3 открытых наборов также не позволило достичь высокого качества диагностики на предлагаемых данных (см. Рис. 20).

Ожидаемо лучше всех себя показало объединение всех 4 базовых наборов (см. Рис. 21).

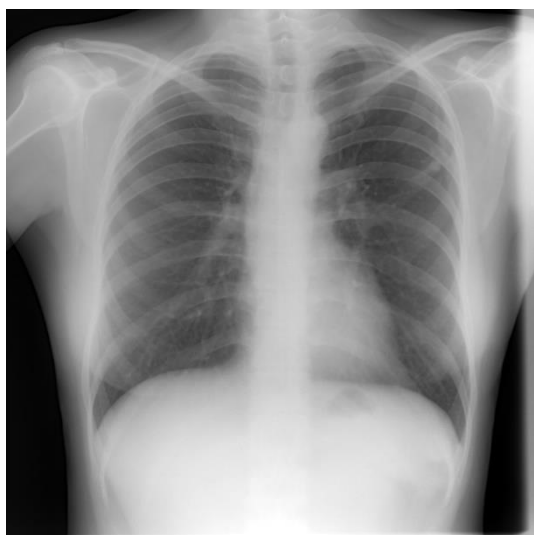
Отметим, что для данной задачи оказалось достаточно эффективным использование алгоритма визуализации результатов анализа. При этом для визуализации области изображения, поддерживающей диагностическое решение, был применён следующий алгоритм, являющийся расширенной модификацией алгоритма Smoothed Score-CAM [20]:

1. Для входного изображения получить фиксированное количество дополнительных изображений с помощью поворота на случайный угол $\varphi \in [-15^\circ, 15^\circ]$ и переноса на случайный вектор (x, y) , $x \in [-30, 30]$, $y \in [-30, 30]$. Полученные изображения вместе с исходным образуют множество M .
2. К каждому изображению I из M применить нейросетевой алгоритм диагностики и сохранить карты активации последнего и предпоследнего свёрточных слоев нейронной сети, а также посчитать веса $\alpha_k^c(I) = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c(I)}{\partial A_{ij}^k}$ с помощью обратного распространения ошибки аналогично весам для алгоритма визуализации Grad-CAM [21], где c – предсказанный алгоритмом диагностики для исходного изображения класс, y^c – оценка уверенности для предсказанного нейронной сетью класса, A^k – карта активации свёрточного слоя в канале k , Z – число пикселей в выбранной карте активации.
3. Следующие шаги выполняются только для тех сохраненных карт активаций, для которых $\alpha_k^c(I) > 0$:
 - а. Карта активаций масштабируется до размера соответствующего ей входного изображения I и умножается на него, образуя новое входное изображение I_{masked} .
 - б. Ко всем изображениям I_{masked} применяется алгоритм диагностики для оценки степени уверенности искомого класса, а затем для каждого канала каждого из рассматриваемых слоёв нейронной сети происходит усреднение полученных оценок по всем изображениям

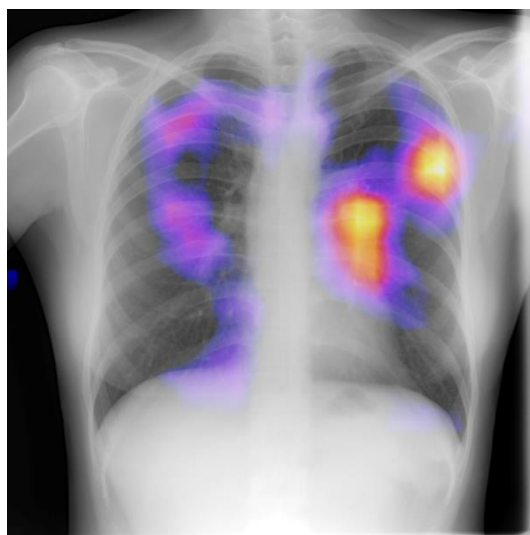
из M . Если после этого усреднённая оценка для искомого класса больше 0.5 для карт активаций с последнего слоя или больше 0.6 для карт активаций с предпоследнего слоя, то такая усреднённая оценка и её карта активаций для исходного изображения сохраняются.

- с. Все сохраненные оценки уверенности искомого класса нормируются функцией $\text{softmax}(\bar{\alpha}^c)$, и отобранные карты активаций суммируются с соответствующими им получившимися весами.

Так, на Рис. 22 приведён пример выделенных алгоритмом областей при положительной диагностике туберкулеза методом EffNetV2-M, обученном на комбинации всех 4 базовых наборов, для снимков из набора Sakha-TB. Подобный подход позволил указать на потенциально важные участки изображения.



а) исходное изображение



б) визуализация области, поддерживающей диагностическое решение

Рис. 22. Пример выходного изображения алгоритма визуализации результатов анализа рентгенограммы

Заключение

В данной работе был представлен Sakha-TB – набор рентгеновских снимков грудной клетки для задачи диагностики туберкулёза, сбалансированный по количеству диагнозов и в некоторой степени сбалансированный по полу и возрасту пациентов.

Было проведено тестирование 2 современных нейросетевых методов классификации изображений на некоторых открытых наборах рентгеновских изображений грудной клетки и на различных комбинациях

рассматриваемых наборов (включая представленный) для решения задачи компьютерной диагностики туберкулёза.

Эксперименты показали, что лишь добавление предлагаемого набора Sakha-TV в обучающую выборку и с некоторым допущением самостоятельное использование набора MC + SZ дают стабильно высокое качество классификации обученным алгоритмом изображений из всех 4 рассмотренных наборов рентгеновских снимков. При этом в случае использования представленного набора в качестве основы не обязательно требуется большой набор изображений. Таким образом, представляется полезным включение набора Sakha-TV в обучающую выборку нейросетевых моделей компьютерной диагностики туберкулёза для повышения стабильности их работы на практике.

Литература

1. *Oloko-Oba M., Viriri S.* A systematic review of deep learning techniques for tuberculosis detection from chest radiograph // *Frontiers in Medicine*. – 2022. – Т. 9. – с. 830515.
2. *Zeyu D., Yaakob R., Azman A.* A review of deep learning-based detection methods for tuberculosis // *2022 IEEE International Conference on Computing (ICOCO)*. – IEEE, 2022. – С. 68–73.
3. *Singh M. et al.* Evolution of machine learning in tuberculosis diagnosis: a review of deep learning-based medical applications // *Electronics*. – 2022. – Т. 11. – № 17. – с. 2634.
4. *Santosh K. C. et al.* Advances in deep learning for tuberculosis screening using chest X-rays: the last 5 years review // *Journal of Medical Systems*. – 2022. – Т. 46. – № 11. – с. 82.
5. *Dovganich A.A. et al.* Automatic out-of-distribution detection methods for improving the deep learning classification of pulmonary X-ray images // *Journal of Image and Graphics (United Kingdom)*. – 2022. – Т. 10. – № 2. – С. 56–63.
6. *Pchelintsev Y.A. et al.* Hardness analysis of X-ray images for neural-network tuberculosis diagnosis // *Computational Mathematics and Modeling*. – 2023. – Т. 33 – № 2. – С. 230–243.
7. *Xue Z. et al.* Cross dataset analysis of domain shift in CXR lung region detection // *Diagnostics*. – 2023. – Т. 13. – № 6. – с. 1068.
8. *Pooch E.H.P., Ballester P., Barros R.C.* Can we trust deep learning based diagnosis? The impact of domain shift in chest radiograph classification // *Thoracic Image Analysis: Second International Workshop, TIA 2020*. – Springer International Publishing, 2020. – С. 74–83.

9. *Candemir S. et al.* Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration // IEEE transactions on medical imaging. – 2013. – T. 33. – № 2. – C. 577–590.
10. *Jaeger S. et al.* Automatic tuberculosis screening using chest radiographs // IEEE transactions on medical imaging. – 2013. – T. 33. – № 2. – C. 233–245.
11. *Chauhan A., Chauhan D., Rout C.* Role of gist and PHOG features in computer-aided diagnosis of tuberculosis without segmentation // PloS one. – 2014. – T. 9. – № 11. – c. e112980.
12. *Liu Y. et al.* Rethinking computer-aided tuberculosis diagnosis // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2020. – C. 2646–2655.
13. *Gaggion N. et al.* Improving anatomical plausibility in medical image segmentation via hybrid graph neural networks: applications to chest x-ray analysis // IEEE Transactions on Medical Imaging. – 2023. – T. 42. – № 2. – C. 546–556.
14. *Larrazabal A.J. et al.* Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis // Proceedings of the National Academy of Sciences. – 2020. – T. 117. – № 23. – C. 12592–12594.
15. *Tan M., Le Q.* Efficientnetv2: Smaller models and faster training // International conference on machine learning. – PMLR, 2021. – C. 10096–10106.
16. *Dosovitskiy A. et al.* An image is worth 16x16 words: Transformers for image recognition at scale // International Conference on Learning Representations. – 2021.
17. *Russakovsky O. et al.* ImageNet large scale visual recognition challenge // International journal of computer vision. – 2015. – T. 115. – C. 211–252.
18. *Brodersen K.H. et al.* The balanced accuracy and its posterior distribution // 2010 20th international conference on pattern recognition. – IEEE, 2010. – C. 3121–3124.
19. *Loshchilov I., Hutter F.* Decoupled weight decay regularization // International Conference on Learning Representations. – 2019.
20. *Wang H. et al.* SS-CAM: Smoothed Score-CAM for sharper visual feature localization // arXiv preprint arXiv:2006.14255. – 2020.
21. *Selvaraju R.R. et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization // Proceedings of the IEEE international conference on computer vision. – 2017. – C. 618–626.