ISSN 1816-0301 (Print)
ISSN 2617-6963 (Online)

## *ОБРАБОТКА ИЗОБРАЖЕНИЙ, РЕЧИ И ТЕКСТА*

## *IMAGE, SPEECH AND TEXT PROCESSING*

УДК 004.9
https://doi.org/10.37661/1816-0301-2020-17-4-48-60

*Поступила в редакцию 17.08.2020*
*Received 17.08.2020*

*Принята к публикации 07.09.2020*
*Accepted 07.09.2020*

# Компьютеризированная диагностика рака простаты на основе полнослайдовых гистологических изображений и методов глубокого обучения

**В. А. Ковалев[1, 2]✉, Д. М. Войнов[2], В. Д. Малышев[1, 2], Е. Д. Лапо[1, 2]**

[1]*Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Беларусь*
✉*E-mail: vassili.kovalev@gmail.com*

[2]*Белорусский государственный университет, Минск, Беларусь*

**Аннотация.** Представлены результаты экспериментальных исследований и разработки средств автоматического анализа и распознавания гистологических изображений с целью получения количественных оценок наличия и степени агрессивности рака простаты в общепринятых шкалах Глисона и ISUP. В качестве исходных данных использовались 10 616 полнослайдовых гистологических изображений с размером большей стороны до 100 000 пикселов и 22 089 их фрагментов размером 256×256 пикселов. Проведена оценка эффективности решения задачи с применением как традиционных методов, так и методов глубокого обучения. В качестве финальных выбраны два решения. Первое решение основано на последовательном анализе фрагментов изображений и включает выделение признаков с использованием сети ResNet50 и последующим обобщением частных результатов распознавания с помощью небольшой сверточной сети. Второе решение базируется на одновременном анализе отобранных информативных участков, представленных в виде промежуточного псевдоизображения, и последующем его распознавании с использованием ансамбля из четырех вариантов сверточных сетей с архитектурой EfficientNetB0. В результате независимого тестирования на закрытом наборе изображений, недоступных авторам, достигнута точность предсказания финальной оценки по шкале ISUP, равная 0,9277.

**Ключевые слова:** рак простаты, гистология, полнослайдовые изображения, глубокое обучение, сверточные нейронные сети

**Благодарности.** Работа была выполнена при частичной финансовой поддержке проекта ГКНТ Беларуси, договор № 225/4/2019. Авторы выражают глубокую благодарность Д. А. Павленко за помощь в поиске и анализе дополнительной информации, необходимой для выполнения данной работы.

**Для цитирования.** Компьютеризированная диагностика рака простаты на основе полнослайдовых гистологических изображений и методов глубокого обучения / В. А. Ковалев [и др.] // Информатика. – 2020. – Т. 17, № 4. – С. 48–60. https://doi.org/10.37661/1816-0301-2020-17-4-48-60

# Computerized diagnosis of prostate cancer based on whole slide histology images and deep learning methods

**Vassili A. Kovalev[1, 2]✉, Dmitry M. Voynov[2], Valery D. Malyshau[1, 2], Elizabeth D. Lapo[1, 2]**

*[1]The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Belarus*
*✉E-mail: vassili.kovalev@gmail.com*

*[2]Belarusian State University, Minsk, Belarus*

**Abstract.** This paper presents the results of an experimental study and the development of tools for automatic analysis and recognition of histological images in order to obtain quantitative estimates of the presence and degree of aggressiveness of prostate cancer in the commonly used Gleason and ISUP scales. The input data consisted of 10 616 whole-slide histological images with the size of the largest side up to 100 000 pixels and 22 089 of their image tiles of 256×256 pixels in size. Two solutions were chosen as the final ones. The first solution is based on sequential analysis of image fragments and includes feature extraction using the ResNet50 network and the subsequent generalization of particular recognition results using a small convolutional network. The second solution is based on the simultaneous analysis of the selected informative sections, presented in the form of an intermediate pseudo-image, and its subsequent recognition using an ensemble of four variants of convolutional networks with the EfficientNetB0 architecture. Being independently tested on an unknown image dataset that was not available for authors, these approaches achieved the prediction accuracy of 0,9277 according to the ISUP scale.

**Keywords:** prostate cancer, histology, whole slide histology, deep learning, convolutional neural networks

**Acknowledgements.** The work was carried out with partial financial support from the project of the State Committee for Science and Technology of Belarus, contract no. 225/4/2019. The authors are also deeply grateful to D. A. Pavlenko for help in finding and analyzing additional information that was necessary for this work.

**For citation.** Kovalev V. A., Voynov D. M., Malyshau V. D., Lapo E. D. Computerized diagnosis of prostate cancer based on whole slide histology images and deep learning methods. *Informatics*, 2020, vol. 17, no. 4, pp. 48–60 (in Russian). https://doi.org/10.37661/1816-0301-2020-17-4-48-60

**Введение.** Рак простаты является одним из распространенных видов онкологических заболеваний у мужчин. Каждый год в мире диагностируется более 1 млн новых случаев заболевания и умирает более 350 тыс. человек [1]. Одним из общепризнанных способов уменьшения смертности от рака простаты является его раннее обнаружение и правильная оценка степени агрессивности. В качестве способа количественной оценки заболевания используется его рейтинг по так называемой шкале Глисона (Gleason Score, [2]), предложенной Дональдом Глисоном. Показатель Глисона определяется визуально по гистологическим изображениям высокого разрешения, которые снимаются с образцов ткани, получаемых в результате биопсии. Существующие варианты данной шкалы незначительно отличаются друг от друга в зависимости от исторически сложившихся предпочтений и некоторых других причин. Однако общей основой оценки является базовая шкала с целыми значениями {0, 3, 4, 5}. При этом уровень 0 соответствует здоровой ткани, уровень 3 назначается участку биопсии с минимальной злокачественностью и положительным прогнозом выживаемости при соответствующем лечении, а уровень 5 обозначает максимальную агрессивность рака с вероятным (иногда уверенным) отрицательным исходом. Соответственно, уровень 4 является промежуточным между 3 и 5. Учитывая то, что изображение образца ткани может быть очень неоднородным как по своим визуальным свойствам, так и по показателям Глисона, в используемой международной системе принято указывать два уровня. Первый уровень соответствует участку ткани, площадь которого является наибольшей на изучаемом образце ткани, а второй – участку, следующему за ним по площади. Указанные номера уровней принято записывать через знак «+». Так, например, образец ткани, бóльшая часть которой представляет собой начальную стадию заболевания, а меньшая – его наиболее агрессивную форму, кодируется как «3+5»; образец, на котором полностью доминирует последняя стадия, представляется в виде «5+5» и т. д. В дальнейшем сводный показатель вычисляется либо просто как сумма указанных чисел, либо, что чаще всего, используется как

некоторый логичный с медицинской точки зрения показатель, определяемый в виде таблично заданной функции типа 0 + 0 = 0, 3 + 3 = 1, 3 + 4 = 2, 4 + 3 = 3, 3 + 5 = 4, 4 + 5 = 5 и 5 + 5 = 5. Указанная интегрированная шкала получила название ISUP (ISUP Score, [2]). Очевидная «нелинейность» шкалы ISUP отражает значительно бóльшую опасность высоких значений, особенно максимального значения показателя Глисона, равного пяти. Это обусловлено тем фактом, что в случае присутствия такой агрессивной ткани, даже в небольшом ее количестве, существует весьма высокая вероятность быстрого развития заболевания в негативном направлении.

С точки зрения машинного обучения в рассматриваемой задаче требуется предсказать некоторое число, которое может иметь одно из нескольких значений шкалы ISUP в интервале 0–5. Поскольку ISUP представляет собой шкалу агрессивности рака и имеет место упорядоченность возможных значений, важно предсказывать не только правильный ответ в виде некоторого абстрактного номера класса, но и значение, наиболее близкое к нему. В первом случае, т. е. без учета наличия отношения порядка на номерах классов, имеет место обычная постановка задачи классификации, в то время как при учете данного факта более корректной является регрессионная постановка.

Вместе с тем следует отметить, что диагностика рака простаты представляет собой достаточно сложную задачу даже для опытных врачей-патологов [1, 3]. Как следствие, в ряде случаев может наблюдаться значительная рассогласованность заключений различных экспертов по одному и тому же пациенту. Сложность диагностики обусловлена целым рядом различных факторов, к числу которых относится большая неоднородность образца ткани, высокая вариабельность морфологического строения и результирующих пространственно-цветовых паттернов даже внутри одного показателя Глисона, существенные различия гистологической картины у разных пациентов, наличие артефактов подготовки, окраски и сканирования образцов ткани и ряд других [4, 5].

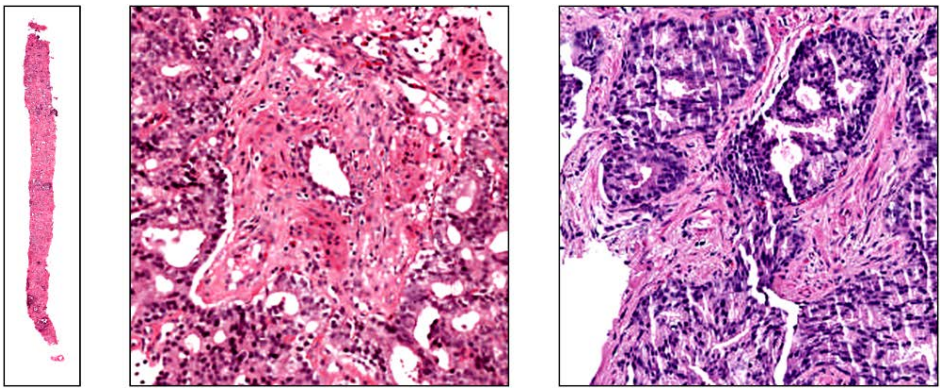
К настоящему времени уже был выполнен ряд работ по проблеме компьютеризированной диагностики и оценки степени злокачественности рака простаты [5–8]. Так, в работах [5, 6] с помощью методов глубокого обучения были получены хорошие результаты при решении задач бинарной классификации ткани простаты (рак против нормы, ранняя стадия рака против поздней). В то же время уровень определения конкретного показателя Глисона был значительно хуже при применении компьютеризированных систем не только в качестве самостоятельного предсказателя ([5, 8], каппа 0,62 и 0,70 соответственно), но и в качестве вспомогательного модуля в процессе поддержки принятия решений гистопатологами ([7], каппа 0,733). При этом во всех рассмотренных работах количество полнослайдовых изображений (ПСИ), использованных для обучения нейронных сетей, было существенно меньше, чем в настоящем исследовании. Тем не менее среди них следует отметить работу [6], в которой исследование проводилось на значительном количестве ПСИ. Однако большинство изображений (11 429) использовалось все же только для тестирования.

Целью данной работы являлось экспериментальное исследование и разработка средств автоматического анализа и распознавания гистологических ПСИ, а также получение общепринятых количественных оценок наличия и степени агрессивности рака простаты для их последующего использования в качестве «второго мнения» врачами-патологами. Работа выполнялась в рамках международного соревнования по диагностике рака простаты PANDA (URL: https://panda.grand-challenge.org/), в котором принимали участие более тысячи команд из разных стран. По результатам независимого тестирования программного обеспечения профильными специалистами на недоступном авторам наборе изображений данная работа заняла 18-е место из 1010 возможных, что соответствует верхней части списка (топ 1,8 %) лауреатов серебряной медали (URL: https://www.kaggle.com/c/prostate-cancer-grade-assessment/leaderboard).

**1. Исходные данные**

***Исходный набор гистологических изображений (табл. 1, акроним DS_RAW).*** Данный набор состоял из 10 616 ПСИ (рис. 1, *а*) с соответствующими показателями в шкалах ISUP и Глисона. Изображения были предоставлены Университетом Неймегена, Голландия (источник А, увеличение х20, пиксел 0,24 мкм, второй уровень с пикселом 0,48 мкм, рис. 1, *б*) и Каролинским институтом, Швеция (источник Б, увеличение х20, пиксел 0,48 мкм, рис. 1, *в*). Следует

отметить, что очевидные цветовые различия изображений обусловлены особенностями технологии подготовки и окраски образцов ткани, характерными для каждого медицинского учреждения, и напрямую не связаны с наличием или отсутствием злокачественных опухолей. Для каждого ПСИ имелись маски, на которых указаны опухолевые области и их показатели Глисона. Из этого набора данных были исключены 671 ПСИ, которые содержали артефакты или у которых информация на маске и показатели не соответствовали друг другу (URL: https://www.kaggle.com/dannellyz/collection-of-600-suspicious-slides-data-loader). В результате очистки был получен набор данных DS_CLEAN. В качестве тестового набора данных использовались два закрытых набора ПСИ. Первый из них включал 420 ПСИ, а второй – 580 ПСИ, доступных только организаторам соревнования.



*а)* *б)* *в)*

Рис. 1. Примеры исходных изображений: *а)* ПСИ размером 18 176×4096 пикселов; *б), в)* тайлы размером 256×256 пикселов из источников А и Б соответственно

Таблица 1

Количественный состав наборов данных

| Акроним | Количество ПСИ | Количество тайлов | Баланс показателей, % | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | ISUP | | | | | | Глисона | | | |
| | | | 0 | 1 | 2 | 3 | 4 | 5 | 0 | 3 | 4 | 5 |
| DS_RAW | 10 616 | – | 27 | 25 | 13 | 12 | 12 | 11 | 18 | 51 | 45 | 13 |
| DS_CLEAN | 9945 | – | 27 | 25 | 13 | 12 | 12 | 11 | 18 | 51 | 45 | 13 |
| DS_SMAL | 3150 | 6000 | 11 | 15 | 15 | 16 | 13 | 30 | 39 | 22 | 24 | 20 |
| DS_MAIN | 5067 | 22 089 | 51 | 32 | – | – | 15 | 2 | 43 | 25 | 19 | 13 |

В процессе разметки входных данных были допущены ошибки, из-за которых до 30 % всех ПСИ имеют неверные показатели в шкалах ISUP и Глисона (URL: https://panda.grand-challenge.org/).

***Предварительный набор изображений (табл. 1, акроним DS_SMAL).*** Все ПСИ на уровне максимального разрешения были разбиты на так называемые тайлы, т. е. некоторые элементарные участки квадратной формы размером 256×256 пикселов. Тайлы, содержащие фон, удалялись. В результате разбиения было получено около 8 млн тайлов. Так как только небольшое количество тайлов относилось к третьему, четвертому и пятому классам Глисона, для получения сбалансированного набора данных, необходимого для обучения нейронных сетей, было выбрано их некоторое подмножество, обеспечивающее равномерное представительство классов. Кроме того, вычисления с использованием полного набора тайлов заняли бы слишком много времени, что делает нецелесообразным использование всех 8 млн. С целью ускорения предварительных вычислительных экспериментов из указанного исходного набора случайным образом было выбрано 6000 тайлов.

***Основной набор изображений (табл. 1, акроним DS_MAIN).*** Данный набор использовался для обучения нейронных сетей, которые определяли оценку по шкале Глисона на уровне одиночного тайла. Во избежание ошибок неверного назначения показателя Глисона из-за наличия

сразу двух его значений у каждого ПСИ нужные тайлы отбирались из ПСИ, которые имели два равных показателя: 3 + 3, 4 + 4, 5 + 5, а также с ПСИ, содержащих нормальную ткань и, соответственно, размеченных как 0 + 0. При использовании такого подхода выявился недостаток тайлов из источника Б с пятым показателем Глисона. Чтобы сбалансировать набор данных, в этот класс были добавлены тайлы из ПСИ с показателями 4 + 5 и 5 + 4 по шкале Глисона. Такой подход уменьшает точность разделения четвертого и пятого классов, но в целом повышает значение тестовой статистической метрики классов каппа [9], вычисляемой следующим образом:

$$\kappa_w = 1 - \frac{q_0}{q_e},$$

$$q_0 = \sum_i \sum_j v_{ij} p_{ij},$$

$$q_e = \sum_i \sum_j v_{ij}\, p_i^* p_j\,,$$

где $p_i^*$ – доля объектов выборки, которые нейронная сеть отнесла к категории $i$ по значениям шкалы ISUP $i$; $p_j$ – доля объектов из категории $j$; $p_{ij}$ – доля объектов, относящихся к категории $j$, которые нейронная сеть предсказала как объекты из категории $i$, $v_{ij} = (i-j)^2$. Все тайлы были выбраны из слоя ПСИ с увеличением, в четыре раза меньшим, чем максимальное, и имели размер 256×256 пикселов. Информация о количественном составе описанных выше наборов изображений представлена в табл. 1.

Следует отметить, что некоторые тайлы, как и целые ПСИ, могут иметь области с различными показателями Глисона, т. е. даже изображения размером 256×256 пикселов могут одновременно содержать участки, представляющие разные классы.

Как следует из табл. 1, в наборе данных DS_MAIN на один ПСИ в среднем приходятся четыре тайла, которые не всегда являются смежными. Таким образом, степень увеличения количества данных как таковых за счет перекрытия тайлов несущественная. Более того, это может внести дополнительные неточности из-за неизбежного увеличения доли фона на тайлах, так как при игольчатой биопсии простаты ширина образца ткани соответствует всего одному-трем тайлам. Отсутствие перекрытия тайлов частично компенсируется их аугментацией, включающей повороты и отражения в процессе обучения нейронных сетей.

**2. Общая схема решения задачи.** Как уже упоминалось ранее, ПСИ могут достигать весьма значительных размеров, вплоть до 100 000 пикселов по каждому измерению с объемом занимаемой памяти до 10 Гб (при обработке без сжатия). При таких размерах использование традиционных методов достаточно затруднительно, а существующие методы глубокого обучения вовсе не приспособлены для анализа данных больших размеров. Кроме того, алгоритмы распараллеливания вычислений на графических ускорителях (GPU) при глубоком обучении не позволяют размещать требуемое количество изображений в графической памяти одновременно даже при использовании самых современных устройств указанного типа. Поэтому для решения данной задачи необходимо проводить определенную предобработку, которая либо сокращает размер исходной задачи (при этом теряется определенная информация), либо делит ее на несколько меньших подзадач, решаемых существующими методами, с последующим объединением частных решений.

Эффективной реализацией *первого подхода* к предобработке является метод, основанный на разбивке исходного изображения по схеме «ПСИ → полотно (мозаика) тайлов». При этом под тайлом здесь понимается некоторый небольшой участок изображения, который содержит фрагмент изучаемой ткани. Сама предобработка включает в себя следующие три шага:

– разделение ПСИ на непересекающиеся тайлы таким образом, чтобы каждый пиксел изображения принадлежал ровно одному тайлу;

– выбор определенным способом среди полученных тайлов заранее заданного числа некоторых представительных тайлов;

– расположение представительных тайлов в прямоугольнике (квадрате) и сохранение их как единого изображения.

При *втором подходе* исходное ПСИ последовательно анализируется и классифицируется тайл за тайлом. По завершении процесса полученные частные результаты распознавания обобщаются с использованием некоторого алгоритма и соответствующий шкале ISUP класс назначается всему ПСИ.

**3. Предварительные эксперименты.** Предварительные эксперименты проводились для оценки сложности решения поставленной задачи распознавания изображений; выявления различий изображений, связанных с их съемкой в разных странах; получения сравнительных оценок ошибок классификации, характерных для разных классов; выяснения представительства и существующих дисбалансов классов в обучающей выборке на уровне отдельных тайлов, а также для других целей, характерных для этапа разведочного анализа данных.

***Оценка эффективности традиционных методов распознавания.*** Оценка эффективности традиционных подходов, включающих вычисление информативных признаков изображений с последующим обучением классификаторов для решения задачи распознавания, выполнялась на усеченном, «пробном» наборе изображений DS_SMAL. Учитывая тот факт, что гистологические изображения представляют собой некоторый специфический вид цветных текстур, в качестве количественных параметров описания изображений использовались следующие известные текстурные признаки:

– текстурные признаки Габора [10] в виде статистических моментов выходных сверток с 40 фильтрами, пятью масштабными уровнями, восемью направлениями;

– гистограммы ориентации градиентов (HOG) с шестью направлениями в интервале 0–180°;

– локальные бинарные шаблоны (LBP) с восемью соседями и радиусами окрестностей $R = 1, 2$;

– улучшенная версия матриц совместной встречаемости цветов пикселов изображений [11], представленных в данном случае в оптимальной палитре из 128 цветов, выбранных с использованием одного из вариантов метода кластеризации на базе K-средних;

– площади и статистические моменты яркостей суперпикселов, получаемых при адаптивном разбиении изображений на суперпикселы методом простой линейной итеративной кластеризации SLIC [12].

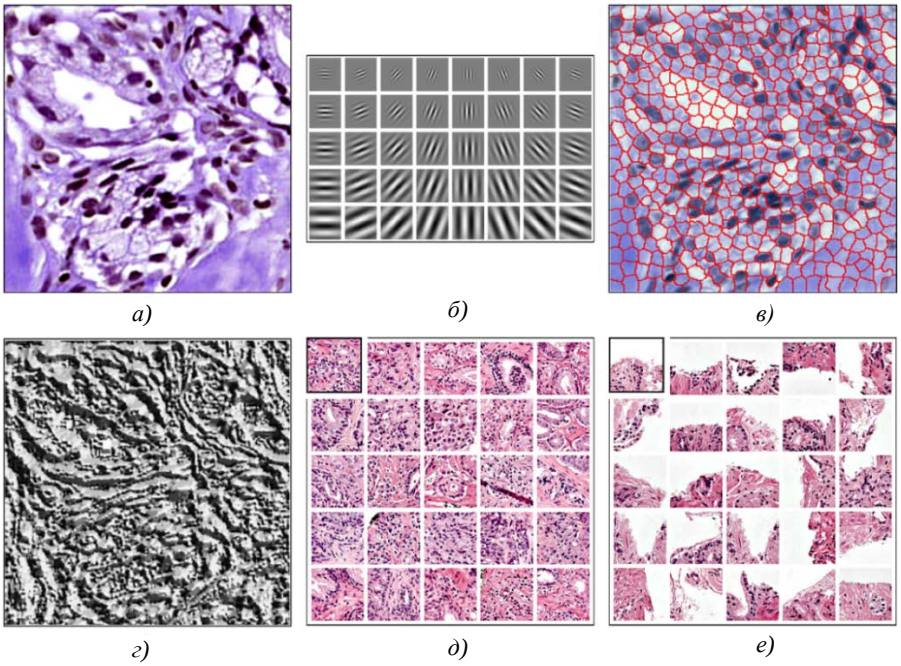
В ситуациях, когда признаки изображений были представлены короткими векторами, например в случае ориентационных гистограмм HOG, они подавались на вход классификаторов напрямую. В случае большого количества элементов дескрипторов изображений, как это имеет место, например, при использовании матриц совместной встречаемости цветов, исходные признаки редуцировались до достаточно малого числа некоррелированных переменных с применением метода главных компонент с порогом отбора 95 % объясняемой вариабельности входных данных. Исключение составляли лишь эксперименты по поиску наиболее похожих изображений по образцу с целью автоматизации процессов формирования обучающих выборок. В этом случае векторизованные версии матриц совместной встречаемости сравнивались напрямую с помощью метрики L1 [13]:

$$\mathrm{L1}(p, q) = \sum_{i=1}^{n} |p_i - q_i|,$$

где $p$, $q$ – векторизованные версии матриц совместной встречаемости. Метод главных компонент со встроенным нормированием переменных также использовался при объединении и последующем совместном использовании разносортных признаков.

В качестве классификаторов применялись такие известные методы, как SVM, Random Forests и kNN. В силу очевидных приоритетов разведочного анализа, мотивированных необходимостью изучения свойств исходных данных как таковых, а не эффективности различных методов классификации и распознавания, другие классификаторы, равно как и оптимизация управляющих параметров указанных классификаторов, не использовались, т. е. опция сеточной грид-оптимизации параметров классификаторов при вызове реализующих их программных модулей была выключена. Все вычислительные эксперименты, рассматриваемые в разд. 3, проводились с использованием языка R, одноименной платформы разработки (URL: https://www.R-project.org/) и входящих в нее библиотек.

Основные типы признаков изображений, перечисленных выше, показаны на рис. 2. Общие результаты предварительных экспериментальных исследований данного направления приведены в табл. 2. Для каждого типа признаков были проведены четыре серии экспериментов по точности предсказания наличия рака и его агрессивности в следующих четырех шкалах: значение ISUP для большего сегмента изображения, указанного в маске; значение Глисона 1; значение Глисона 2; значение в шкале ISUP. Кроме точного предсказания в виде конкретного значения класса, задача предсказания также решалась для некоторых объединений соседних классов, что имеет определенный медицинский смысл. Например, рак отсутствует (0), рак умеренно агрессивен (3∪4) или это крайне агрессивная форма рака (5). Поскольку точность предсказания в зависимости от типа используемых классификаторов варьировала незначительно (порядка 0,5–1,5 %), тип классификатора в табл. 2 не указывается.



*а)* *б)* *в)* *г)* *д)* *е)*

Рис. 2. Исходное тестовое гистологическое изображение (тайл 256×256 пикселов) (*а*); используемые фильтры Габора (*б*); суперпиксельное представление исходного изображения (*в*); карта локальных бинарных шаблонов (*г*); примеры поиска изображений по двум образцам, представленным в левых верхних углах с применением матриц совместной встречаемости цветов (*д*, *е*)

Таблица 2

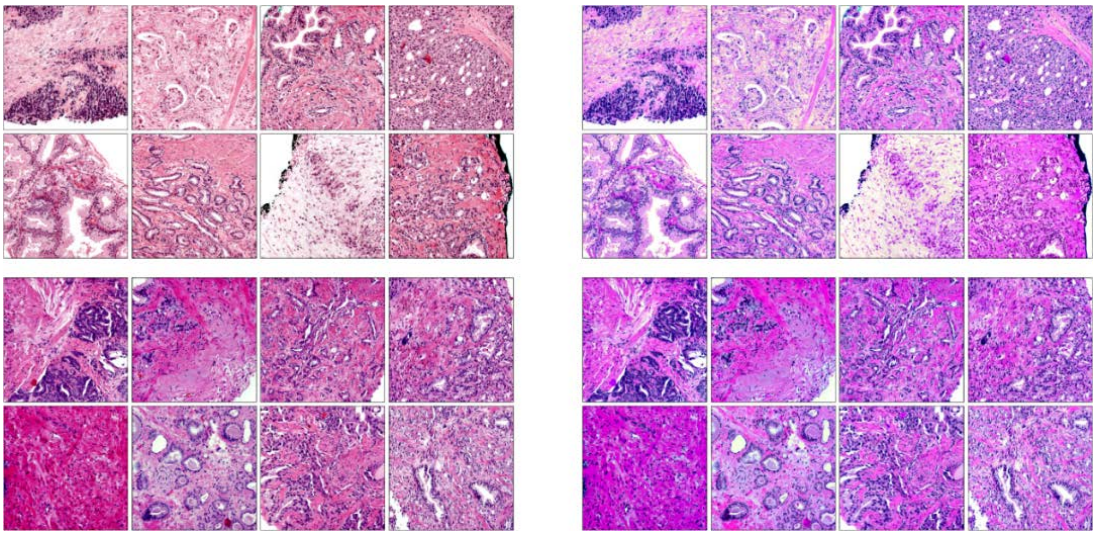
Точность предсказания опасности рака традиционными методами

| Тип признаков, схема объединения классов | Тип шкалы, количество классов | | | |
|---|---|---|---|---|
| | Больший сегмент, пять классов | Глисон 1, четыре класса | Глисон 2, четыре класса | ISUP, пять классов |
| Габор и HOG | 0,418 | 0,493 | 0,506 | 0,491 |
| Суперпикселы | 0,435 | 0,532 | 0,534 | 0,528 |
| Габор и суперпикселы | 0,448 | 0,539 | 0,534 | 0,532 |
| Совместная встречаемость цветов | 0,452 | 0,535 | 0,531 | 0,530 |
| Все признаки | 0,466 | 0,550 | 0,545 | 0,539 |
| Все признаки, (0∪1∪2), (3∪4), (5) | 0,597 | – | – | 0,603 |
| Все признаки, (0), (3∪4), (5) | – | 0,603 | 0,611 | – |

Представленные в табл. 2 результаты показывают, что при использовании традиционных методов точность предсказания наличия и степени агрессивности рака достаточно невысокая. Вероятнее всего, причинами этого является целый ряд факторов, включая ориентацию каждого типа традиционных признаков на конкретные, наперед заданные свойства изображений; отсутствие механизма адаптации к особенностям изображений, заданных обучающей выборкой;

слабую способность традиционных классификаторов к обобщению пространства признаков (generalization); малую «емкость» набора обучаемых параметров (десятки и сотни против сотен тысяч и миллионов у сверточных нейронных сетей) и др. Кроме того, при интерпретации результатов предсказания, приведенных в табл. 2, следует учитывать, что разметка классов в пробном наборе изображений DS_SMAL сильно зашумлена, т. е. количество ошибок назначения классов медиками точно неизвестно.

***Предварительные эксперименты с использованием методов глубокого обучения.*** Для решения задачи классификации тайлов был обучен ряд общедоступных архитектур сверточных сетей, включая ResNet50, MobileNetV2, VGG16. Процесс обучения сетей выполнялся с применением оптимизатора RMSprop. Когда потери на валидационной выборке изображений переставали уменьшаться, скорость обучения (learning rate) снижалась в 10 раз. К входным данным применялась аугментация, включающая в себя стандартные геометрические операции, такие как отражение, случайный поворот на 90° один или более раз, отражение по диагонали. В связи с тем что изображения из представленных источников имеют различные цветовые палитры, к ним была применена цветовая нормализация методом Масенко [14] с использованием маски, основанная на декомпозиции по цвету. Нормализованные тайлы из источников данных А и Б показаны на рис. 3. В обоих случаях верхние восемь тайлов получены из источника А, а нижние восемь – из источника Б. В ходе тестирования наилучшие результаты были получены с помощью архитектуры ResNet50, поэтому она была выбрана для всех последующих экспериментов. Результаты проведенных экспериментов приведены в табл. 3.



*а)* *б)*

Рис. 3. Примеры исходных изображений (*а*) и их нормализованные версии (*б*)

Таблица 3

Результаты предварительных экспериментов по распознаванию тайлов

| Архитектура сети, дополнительные условия | Точность классификации | Взвешенная каппа-метрика |
|---|---|---|
| ResNet50 | 0,87 | 0,87 |
| MobileNetV2 | 0,82 | 0,81 |
| VGG16 | 0,68 | 0,71 |
| ResNet50 (с MaxPooling) | 0,84 | 0,83 |
| ResNet50 (с удалением ошибочных тайлов) | 0,86 | 0,83 |

В качестве эксперимента также были сделаны попытки обучения сети с каппа-потерями (kappa-loss) в качестве функции потерь, однако такой метод не позволил получить значение по метрике «точность» (accuracy) выше 0,59. Вариант с усреднением категориальной функции потерь и каппа-потерями оказался тоже неэффективным.

В связи с тем что сеть, обученная для классификации тайлов, не использовалась явно для предсказания, а выступала в роли алгоритма извлечения признаков изображений (feature extractor), с целью потенциального улучшения представления получаемой карты признаков была обучена нейронная сеть, основанная на архитектуре ResNet50, где предпоследний слой GlobalAveragePooling был заменен на MaxPooling. Данная модель не превзошла рассмотренные на этапе классификации тайлов модели и не дала видимых преимуществ на последующих этапах решения задачи.

При детальном изучении результатов тестирования было отмечено, что среди тестовых тайлов, вырезанных из ПСИ, которые получены из источника Б, есть достаточно большой процент тех, что были помечены как раковые, но ошибочно классифицировались сетью как норма. Этот результат мог получиться за счет того, что для изображений из источника Б была предоставлена только приблизительная область, в которой находится опухоль, однако точная локализация раковых клеток не была выполнена. Следовательно, при выделении тайлов из такой области некоторые из них могли действительно не содержать образцов ткани с раком. Для очистки данных от подобных тайлов применялась следующая последовательность действий:

1) обучение сети на всех данных;
2) предсказание обучающей и валидационной выборки с помощью обученной сети;
3) исключение из выборок тайлов с показателем Глисона 3, 4, 5, принадлежащих источнику Б, для которых была предсказана оценка по шкале Глисона 0;
4) обучение сети с использованием новых данных.

Представленный подход позволил повысить значение метрики «отзыв» (recall) для показателя Глисона 0, однако значение данной метрики для остальных оценок по шкале Глисона упало.

Реализация всех методов решения задачи с помощью второго подхода (см. разд. 2), основанного на классификации изображений с помощью нейронных сетей, была осуществлена на языке программирования Python с использованием библиотеки Keras [15], которая позволяет интегрировать такие упомянутые выше архитектуры нейронных сетей, как ResNet50, MobileNetV2, VGG16.

Предварительные эксперименты также проводились для отработки первого подхода к решению задачи (см. разд. 2), который предполагал формирование промежуточного изображения-полотна из представительных тайлов. В качестве представительных выбирались тайлы, которые содержали наибольшую долю (площадь), занимаемую непосредственно тканью простаты. Примеры подобных участков представлены на рис. 2, *д*, в то время как неподходящие тайлы с малой площадью ткани – на рис. 2, *е*. Такой способ показал хорошую эффективность ввиду специфики анализируемых данных: исходные ПСИ содержали снимки тонкой полоски ткани простаты, а большую часть изображения занимал фон.

В результате предобработки описанным способом каждому ПСИ из исходного набора ставится в соответствие гораздо меньшее изображение, которое тем не менее является достаточно представительным в отношении разнообразия участков ткани. Благодаря этому набор данных, состоящий из пар типа «ПСИ – ISUP score», преобразуется в набор пар типа «полотно тайлов – ISUP score», в котором полотно из тайлов можно подвергнуть анализу существующими методами глубокого обучения. Учитывая наличие некоторой упорядоченности значений шкалы ISUP, задача диагностики рака простаты представлялась как задача ординарной регрессии. Для этого метка каждого изображения отображалась в вектор из пяти чисел так, что для метки $M$ в полученном векторе первые $M$ элементов принимали значение 1, а остальные элементы – значение 0. Например, значение 0 представлялось как [0, 0, 0, 0, 0], а значение 3 – как [1, 1, 1, 0, 0].

Полученная задача решалась известными методами глубокого обучения, а именно с помощью глубокой нейронной сети в роли предсказателя. В качестве архитектуры нейронной сети была выбрана EfficientNet-B0 ввиду ее высокой эффективности при достаточно малом размере и относительно небольшом времени, необходимом для ее обучения. С учетом того что была поставлена задача ординарной регрессии, построенная нейросетевая модель обладала следующими особенностями:

– последний слой сети состоял из пяти нейронов;

– для того чтобы элементы выходного вектора находились в интервале от 0 до 1, в качестве функции активации была выбрана логистическая функция (сигмоид);

– для получения значения шкалы ISUP выход такой сети суммировался и округлялся. Например, вектор предсказаний [0,9, 0,7, 0,8, 0,1, 0,2] соответствовал значению 3 шкалы ISUP.

С учетом приведенных выше особенностей в качестве функции потерь была выбрана бинарная кросс-энтропия [16]. В случае, когда предсказание идет для некоторых двух классов А и Б, кросс-энтропию можно вычислить по формуле

$$L_{\text{cross-entropy}} = -(y \log(p) + (1 - y) \log(1 - p)),$$

где $y$ – бинарный индикатор 0 или 1 в зависимости от того, является ли метка класса А правильной классификацией для текущего наблюдения, а $p$ – предсказанная вероятность того, что текущее наблюдение принадлежит классу А.

Таблица 4

Результаты предсказания значения шкалы ISUP по изображению-полотну

| Входной набор данных | Размер полотна тайлов | Взвешенная каппа-метрика |
|---|---|---|
| 1. DS_CLEAN | 4×4 | 0,884 |
| 2. DS_CLEAN | 5×5 | 0,892 |
| 3. DS_CLEAN | 6×6 | 0,881 |
| 4. DS_RAW | 5×5 | 0,883 |
| 5. DS_RAW | 6×6 | 0,883 |

Из табл. 4 видно, что изменение размеров изображения-полотна несущественно влияло на качество конечных результатов распознавания. Кроме того, проводились эксперименты с решением исходной задачи в классификационной, а не регрессионной постановке. Однако качество получаемых результатов в этом случае было существенно ниже.

**4. Финальные эксперименты**

***Решение, основанное на последовательной классификации тайлов ПСИ.*** В данном случае задача решалась согласно второму подходу в два этапа. На первом этапе проводился последовательный анализ тайлов, далее на втором этапе полученные результаты анализа обобщались для всего ПСИ. Оба этапа выполнялись с помощью соответствующих нейронных сетей. На первом этапе использовалась сеть, основанная на архитектуре ResNet50 (см. разд. 3). Результатом ее работы были признаки, необходимые для предска-зания показателей Глисона для каждого текущего тайла анализируемого ПСИ. На втором этапе применялась некоторая небольшая сверточная нейронная сеть, входом которой являлись упомянутые признаки показателей Глисона, а выходом – финальные показатели в шкале ISUP. Выполнение подхода было обусловлено тем, что в существующей медицинской практике показатель ISUP определяется на основе соотношения областей на ПСИ с различными показателями Глисона.

В техническом плане в качестве входных признаков второй, обобщающей нейронной сети были выбраны 2048 признаков с предпоследнего слоя первой сети. Из этих признаков формировалось некоторое псевдоизображение. Таким образом, каждому ПСИ соответствовало псевдоизображение размером $WT \times HT \times 2048$, где $WT$ и $HT$ – это количество тайлов в ширину и высоту ПСИ соответственно. Именно такие псевдоизображения подавались на вход второй сверточной нейронной сети для получения выходного показателя ISUP.

Поскольку пространственное разрешение псевдоизображений, получаемых после работы первой сети, было слишком мало, использовать на втором этапе все преимущества и обобщающие возможности таких широко известных нейронных сетей, как VGG16, ResNet, EfficientNet и др., не представлялось возможным. Поэтому была предложена некоторая небольшая сеть, архитектура которой детально представлена в табл. 5. В этой нейронной сети между всеми сверточными слоями использовалась функция ReLU и пакетная нормализация.

Для предсказания показателя ISUP с помощью второй нейронной сети применялись такие функции потерь, как категориальная кросс-энтропия, каппа-функция или их комбинация.

Таблица 5

Сверточная сеть, используемая для предсказания показателей ISUP

| Тип слоя | Количество каналов | |
|---|---|---|
| | входных | выходных |
| Сверточный слой 3×3×2 + MaxPool | 2048 | 1024 |
| Сверточный слой 3×3×2 + MaxPool | 1024 | 512 |
| Сверточный слой 3×3×2 + MaxPool | 512 | 256 |
| Сверточный слой 3×3×2 + MaxPool | 256 | 128 |
| Полносвязный слой | 128 | 32 |
| Полносвязный слой | 32 | 16 |
| Полносвязный слой | 16 | 6 |

Поскольку категориальная кросс-энтропия не учитывает порядок в распределении классов, с ее помощью не удалось добиться высоких показателей тестовой метрики. С другой стороны, каппа-функция в качестве функции потерь является довольно чувствительной к выбросам, что не позволило добиться эффективного процесса обучения. Поэтому было принято решение использовать комбинированную функцию, которая представляет собой сумму категориальной кросс-энтропии, и каппа-функцию. В процессе обучения энтропия быстро получает оптимальное значение, после чего нейронная сеть оптимизирует в основном каппа-функцию, что позволяет достичь лучшие показатели точности.

В результате реализации описанной методики были получены следующие значения точности предсказания показателей ISUP ПСИ: 0,807 на доступной тестовой выборке и 0,811 на неизвестной авторам тестовой выборке.

***Решение, основанное на промежуточном изображении из информативных тайлов.*** Задача решалась согласно первому подходу, т. е. методом формирования промежуточного изображения-полотна из информативных тайлов. В качестве последних выбирались тайлы, которые содержали наибольшую долю (площадь), занимаемую непосредственно тканью простаты. В отличие от предыдущего решения в данном случае задача решалась не в классификационной, а регрессионной постановке. Для повышения качества распознавания использовались различные приемы, основные из них представлены ниже:

1. Построение ансамбля моделей. Среди имеющихся обученных нейронных сетей выбиралось некоторое их подмножество. Каждая модель из подмножества совершала предсказание значений шкалы ISUP каждого изображения из тестовой выборки, в качестве ответа выбиралось среднее значение их предсказаний. Стоит отметить, что существуют более сложные методы построения ансамблей моделей, такие, например, как взвешенное среднее, обучение модели по предсказаниям и др. Однако в рамках данной работы выбранный способ показал достаточно большой прирост качества предсказания наличия рака и его стадию.

2. Сдвинутая нарезка тайлов. Помимо нарезки тайлов с верхнего левого угла ПСИ также проводилась нарезка с точки, сдвинутой от этого угла на 64, 128, 192 пиксела вниз и влево. После такой нарезки конструировалось новое изображение-полотно и процесс предсказания выполнялся заново. В результате получались четыре различных предсказания, из которых затем вычислялось среднее. Этот подход позволяет найти различные текстурные особенности изображения, которые могли остаться незамеченными в результате оригинальной нарезки.

3. Аугментация тестовой выборки. В последнее время данный метод стабилизации предсказаний получил очень широкое распространение. Суть его заключается в проведении различных трансформаций изображений и предсказания как для исходной тестовой выборки, так и для трансформированной с последующим усреднением результатов. Назначение метода аналогично аугментации при обучении – это ослабление эффекта переобучения под какие-либо визуальные особенности изображений, не имеющие значения при решении задачи.

В окончательном варианте был применен прием ансамблирования. Среди пяти обученных нейронных сетей, представленных в табл. 4, для построения финального ансамбля были выбра-

ны обученные сети 1, 3, 4 и 5. Сеть 2 не выбиралась по причине более низкого качества ансамблей, получившихся при ее использовании. В качестве метода аугментации был выбран поворот изображений на 90°.

Решение, построенное по описанной методике, получило следующие значения взвешенной каппа-метрики предсказания финальных значений по шкале ISUP: 0,8863 на публичной тестовой выборке и 0,9277 на закрытой выборке, недоступной авторам. Стоит отметить, что аналогичное решение, в котором для ансамбля выбирались только сети 1 и 3, обученные на очищенной выборке, показало значения качества 0,8836 и 0,9297 соответственно. Однако это было обнаружено уже после завершения упомянутого выше международного соревнования по диагностике рака простаты и поэтому данные результаты официально приняты не были.

**Заключение.** Методы распознавания изображений, основанные на глубоком обучении, обеспечивают существенно более высокое качество решения задач распознавания изображений по сравнению с традиционными подходами, основанными на вычислении признаков (feature extraction) и последующем применении известных классификаторов типа SVM, Random Forests, kNN и др.

Подход, основанный на последовательном распознавании фрагментов изображений (тайлов) с последующим обобщением результатов, показал относительно низкое качество решения задачи и оказался неконкурентоспособным по сравнению с альтернативным подходом.

Наличие определенных отношений порядка типа лучше-хуже на выходных классах должно быть учтено при постановке задачи распознавания и ее реализации в виде нейронной сети, сконфигурированной для работы в классификационной или регрессионной постановке.

С учетом специфики исходных данных замена полнослайдовых изображений на подмножество представительных тайлов позволила построить нейросетевые модели высокого качества, несмотря на существенную потерю информации.

Применяемые в настоящей работе техники стабилизации предсказаний типа ансамблей нейронных сетей и аугментации тестовых данных давали большой прирост качества, в то время как тренировка более сложных нейросетевых моделей чаще всего заканчивалась их переобучением.

## References

1. Rawla P. Epidemiology of prostate cancer. *World Journal of Oncology*, 2019, vol. 10, no. 2, pp. 63–89.

2. Epstein J. I, Allsbrook W. C. Jr, Amin M. B., Egevad L. L. ISUP Grading Committee. The 2005 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma. *The American Journal of Surgical Pathology*, 2005, vol. 29, iss. 9, pp. 1228–1242.

3. Camparo P., Egevad L., Algaba F., Berney D. M., Boccon-Gibod L., …, Varma M. Utility of whole slide imaging and virtual microscopy in prostate pathology. *Acta Pathologica, Microbiologica, et Immunologica Scandinavica*, 2012, vol. 120, iss. 4, pp. 298–304.

4. Goldenberg S. L., Nir G., Salcudean S. E. A new era: artificial intelligence and machine learning in prostate cancer. *Nature Reviews Urology*, 2019, vol. 16, pp. 391–403.

5. Ström P., Kartasalo K., Olsson H., Solorzano L., Delahunt B., …, Eklund M. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *The Lancet Oncology*, 2020, vol. 21, iss. 2, pp. 222–232.

6. Pantanowitz L., Quiroga-Garza G., Bien L., Heled R., Laifenfeld D., …, Dhir R. An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. *The Lancet Digital Health*, 2020, vol. 2, iss. 8, pp. e407–e416.

7. Bulten W., Balkenhol M., Belinga J.-J. A., Brilhante A., Çakır A., …, Litjens G. Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. *Modern Pathology*, 2020. Available at: https://arxiv.org/abs/2002.04500 (accessed 06.08.2020).

8. Nagpal K., Foote D., Liu Y., Chen P.-H. C., Wulczyn E., …, Stumpe M. C. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *Nature Partner Journal Digital Medicine*, 2019, vol. 2, iss. 48, pp. 1–10.

9. Schuster C. A note on the interpretation of Weighted Kappa and its relations to other rater agreement statistics for metric scales. *Educational and Psychological Measurement*, April 2004, vol. 64, no. 2, pp. 243–253.

10. Luan S., Chen C., Zhang B., Han J., Liu J. Gabor convolutional networks. *IEEE Transactions on Image Processing*, 2018, vol. 27, no. 9, pp. 4357–4366.

11. Kovalev V., Volmer S. Color co-occurrence descriptors for querying-by-example. *International Conference on Multimedia Modeling, Lausanne, Switzerland, 12–15 October 1998*. Lausanne, 1998, pp. 32–38.

12. Achanta R., Shaji A., Smith R., Lucchi A., Fua P., Susstrunk S. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on PAMI*, 2012, vol. 34, no. 11, pp. 2274–2282.

13. Horn R. A., Johnson C. R. Matrix Analysis. Part 5. *Norms for Vectors and Matrices*. England, Cambridge University Press, 1990.

14. Macenko M., Niethammer M., Marron J., Borland D., Woosley J. T., …, Thomas N. E. A method for normalizing histology slides for quantitative analysis. *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Boston, MA, USA, 28 June – 1 July 2009*. Boston, 2009, pp. 1107–1110.

15. Gulli A., Sujit P. *Deep learning with Keras*. Packt Publishing Ltd, 2017, 318 p.

16. Rubinstein R. Y., Kroese D. P. *The Cross Entropy Method: a Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation (Information Science and Statistics)*. Berlin, Heidelberg, Springer-Verlag, 2004, 321 p.

## Информация об авторах

*Ковалев Василий Алексеевич*, кандидат технических наук, заведующий лабораторией анализа биомедицинских изображений, Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Беларусь, доцент, Белорусский государственный университет, Минск, Беларусь.
E-mail: vassili.kovalev@gmail.com

*Войнов Дмитрий Михайлович*, магистрант, Белорусский государственный университет, Минск, Беларусь.
E-mail: voynovdd@gmail.com

*Малышев Валерий Дмитриевич*, инженер-программист лаборатории анализа биомедицинских изображений, Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Беларусь, магистрант, Белорусский государственный университет, Минск, Беларусь.
E-mail: malyshevalery@gmail.com

*Лапо Елизавета Дмитриевна*, инженер-программист лаборатории анализа биомедицинских изображений, Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Беларусь, магистрант, Белорусский государственный университет, Минск, Беларусь.
E-mail: lilibetlapo@gmail.com

## Information about the authors

*Vassili A. Kovalev*, Cand. Sci. (Eng.), Head of the Laboratory of Biomedical Images Analysis, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Belarus, Associated Professor, Belarusian State University, Minsk, Belarus.
E-mail: vassili.kovalev@gmail.com

*Dmitry M. Voynov*, Undergraduate, Belarusian State University, Minsk, Belarus.
E-mail: voynovdd@gmail.com

*Valery D. Malyshau*, Software Engineer of the Laboratory of Biomedical Image Analysis, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Belarus, Undergraduate, Belarusian State University, Minsk, Belarus.
E-mail: malyshevalery@gmail.com

*Elizabeth D. Lapo*, Software Engineer of the Laboratory of Biomedical Image Analysis, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Belarus, Undergraduate, Belarusian State University, Minsk, Belarus.
E-mail: lilibetlapo@gmail.com