DOI: 10.15690/vramn946

И.А. Гундырев[1], Л.В. Бельская[2, 3], В.К. Косенок[3, 4], Е.А. Сарф[3]

[1] ООО «Три-Софт», Омск, Российская Федерация
[2] Омский государственный технический университет, Омск, Российская Федерация
[3] ООО «ХимСервис», Москва, Российская Федерация
[4] Омский государственный медицинский университет, Омск, Российская Федерация

Применение синтетических образов для решения задачи классификации на примере диагностики рака легкого

Обоснование. *С математической точки зрения задачи медицинской диагностики представляют собой задачи классификации данных. При этом важно понимать, насколько существенные искажения могут внести в результат классификации погрешности сбора первичной диагностической информации, в частности результатов биохимических тестов.* ***Цель исследования*** *— установление зависимости результата классификации от вариативности первичной диагностической информации на примере модельного классификатора.* ***Методы***. *В исследовании случай-контроль приняли участие пациенты, которые были разделены на 2 группы — основную (с диагнозом рака легкого, n=200) и контрольную (условно здоровые, n=500). Всем участникам было проведено биохимическое исследование слюны, а также последующая гистологическая верификация диагноза. Биохимический состав слюны определен спектрофотометрически. На основе полученных данных построен модельный классификатор для диагностики рака легкого (случайный лес). В каждый параметр, лежащий в основе классификатора, вносили отклонения в заданном диапазоне (±1–5%, ±5–10%, ±10–15%), создавая синтетические образы. Затем методом кросс-валидации проведена оценка результатов классификации.* ***Результаты***. *Определены базовые диагностические характеристики модельного классификатора (чувствительность — 72,5%; специфичность — 86,0%). При увеличении отклонений синтетических образов от базового значения диагностические характеристики при общей классификации ухудшаются. Однако результат уверенной классификации, напротив, дает более высокие значения (чувствительность — 81,8%, специфичность — 93,1%). В случае уверенной классификации близкие образы, которые по результатам классификации попадают в разные классы, удаляются, тогда как в случае общей — учитываются. Разница между методами классификации связана с наличием образов, на которых классификатор дает результат принадлежности к классу в диапазоне 0,45–0,55. Поэтому необходимо введение третьего класса в классификатор, так называемой серой зоны (0,4–0,6), т.к. вероятность постановки ошибочного диагноза в данной области существенно повышается.* ***Заключение***. *Полученные результаты позволяют сделать вывод, что измерительная погрешность в диапазоне (±1–15%) не оказывает существенного влияния на качество классификации.*

Ключевые слова: *слюна, рак легкого, классификатор, случайный лес, кросс-валидация.*

(***Для цитирования***: Гундырев И.А., Бельская Л.В., Косенок В.К., Сарф Е.А. Применение синтетических образов для решения задачи классификации на примере диагностики рака легкого. *Вестник РАМН*. 2018;73 (2):96–104. doi: 10.15690/vramn946)

Обоснование

В настоящее время не вызывает сомнений актуальность применения математических методов в медико-биологических исследованиях [1]. При этом особо перспективными являются методы многомерного статистического анализа, которые позволяют как систематизировать и обрабатывать результаты медицинских исследований, так и выявлять структуру взаимосвязей между отдельными признаками [2]. С математической точки зрения задача медицинской диагностики представляется задачей классификации [3–5]. Для решения подобных задач широко применяются методы машинного обучения [6–9]. Один из способов машинного обучения — обучение с учителем (*Supervised learning*) — используется для создания классификаторов по множеству так называемых базовых образов, т.е. результатов анализов конкретных пациентов, принадлежность которых к классу «Здоров/Болен» известна (например, диагноз подтвержден гистологически). На следующем этапе ставится задача спрогнозировать принадлежность нового образа (нового пациента) определенному классу, т.е. поставить предварительный диагноз.

Однако исследования в медицине усложняются необходимостью сбора значительного количества первичных данных, получаемых, как правило, методами клинической лабораторной диагностики [10]. В основу классификатора вносятся данные, выраженные в числовой форме. При этом возникает несколько вопросов: во-первых, насколько существенные расхождения могут быть получены при использовании одного и того же классификатора с данными, полученными в разных лабораториях, на разных приборах и т.д., во-вторых, какое количество базовых (обучающих) образов оптимально для получения надежного диагностического результата. Для каждого базового образа (результатов анализов конкретного пациента) будем рассматривать образы с малыми отклонениями в каждом параметре как единый объект. Такие образы легко получить синтетически, а именно внесением в каждый параметр исходного образа отклонения в заданном диапазоне (±1–5%, ±5–10%, ±10–15%). Из каждого базового образа создаются синтетические образы — n=10, n=30, первый из которых совпадает с базовым. При этом образы, полученные из одного базового, называются близкими. Необходимо отметить, что получение базовых (обучающих) образов для медицинской диагностики — процесс, затратный и по времени, и по ресурсам лаборатории. Но именно разнообразие таких образов обеспечивает лучшее отражение в объектах, известных классификатору, всей генеральной совокупности пациентов.

В качестве удобной модели для построения классификатора выбран разработанный нами алгоритм диагностики рака легкого [11]. В основу модельного классификатора

положено одновременное определение в слюне восьми биохимических параметров, таких как сиаловые кислоты [12–16], производные гистидина (гистамин, уроканиновая кислота) [17–20], общий белок [21], мочевина [22], активность ряда ферментов (щелочная фосфатаза, каталаза) [23, 24], среднемолекулярные токсины [25, 26]. Данный алгоритм не соответствует клиническим рекомендациям по диагностике рака легкого, однако в его основе лежит многомерная статистическая обработка данных, что позволяет показать возможности применения синтетических образов для решения задач классификации, которые могут быть впоследствии использованы при работе с любым другим алгоритмом.

Цель исследования — установление зависимости результата классификации от вариативности первичной диагностической информации на примере модельного классификатора.

Методы

Дизайн исследования

В исследовании случай-контроль приняли участие пациенты, которые были разделены на 2 группы — основную (с диагнозом рака легкого) и контрольную (условно здоровые). Включение в группы происходило параллельно. Всем участникам проведено биохимическое исследование слюны.

Критерии соответствия

В качестве *критериев включения* рассматривались мужской пол, возраст пациентов 30–75 лет; отсутствие какого-либо лечения на момент забора слюны (в день госпитализации), в том числе хирургического, химиотерапевтического или лучевого; отсутствие признаков активной инфекции (включая гнойные процессы); проведение санации полости рта; гистологически подтвержденный диагноз рака легкого независимо от типа и стадии.

Критерии исключения: отсутствие гистологической верификации диагноза.

Условия проведения

Пациенты основной группы были обследованы в Клиническом онкологическом диспансере (Омск, Российская Федерация). Пациенты для контрольной группы были набраны в рамках проведения плановой диспансеризации на базе городской поликлиники № 4 (Омск, Российская Федерация). Биохимические исследования осуществляли в лаборатории ООО «ХимСервис» (Омск, Российская Федерация).

Продолжительность исследования

Исследование проводилось в период 2016–2017 гг.

Описание медицинского вмешательства

У всех участников до начала лечения проводили забор слюны в количестве 2 мл. Образцы слюны собирали утром натощак путем сплевывания в стерильные пробирки, центрифугировали при 7000 об./мин. Пациентам контрольной группы было проведено флюорографическое обследование. Пациенты основной группы были госпитализированы для радикальной операции в объеме лобэктомии, билобэктомии, пневмонэктомии, комбинированного лечения или видеоторакоскопии для

I.A. Gundyrev[1], L.V. Bel'skaya[2, 3], V.K. Kosenok[3, 4], E.A. Sarf[3]

[1] Tri-Soft, Omsk, Russian Federation
[2] Omsk State Technical University, Omsk, Russian Federation
[3] Omsk State Medical University, Omsk, Russian Federation
[4] ChemService, Moscow, Russian Federation

The Use of Synthetic Images for Solving the Classification Problem by the Example of Lung Cancer Diagnosis

***Background**: From a mathematical point of view, the problems of medical diagnostics are the tasks of data classification. It is important to understand how significant distortions can contribute to the result of classification errors in the collection of primary diagnostic information, in particular, the results of biochemical tests. **Aims**: Determination of the dependence of the prediction result on the variability of the primary diagnostic information on the example of the model classifier. **Materials and methods**: The case-control study enrolled patients who were divided into 2 groups: the main (diagnosed with lung cancer, n=200) and the control group (conditionally healthy, n=500). Questioning and biochemical saliva study was performed in all participants. Patients of the main group and the comparison group were hospitalized for surgical treatment, after which carried out the histological verification of the diagnosis. The biochemical composition of saliva is determined spectrophotometrically. Based on the data obtained, a model classifier for the diagnosis of lung cancer (a random forest) has been constructed. In each parameter underlying the classifier, deviations were made in the specified range (±1–5%, ±5–10%, ±10–15%), creating synthetic images. Then, the results of the classification were evaluated by the cross-validation method. **Results**: The basic diagnostic characteristics of the model classifier are determined (sensitivity — 72.5%, specificity — 86.0%). As the deviations of synthetic images from the baseline increase, diagnostic characteristics deteriorate with the general classification. However, the result of a confident classification, on the contrary, gives higher values (sensitivity — 81.8%, specificity — 93.1%). In case of a confident classification, similar images that fall into different classes according to the classification results are deleted, whereas in the case of a general classification, they are taken into account. The difference between methods of classification is associated with the presence of images on which the classifier gives the result of belonging to the class in the range of 0.45–0.55. Therefore, it is necessary to introduce a third class into the classifier, the so-called gray zone (0.4–0.6), since the probability of making an erroneous diagnosis in this area is significantly increased. **Conclusions**: The obtained results allow to conclude that the measurement error in the range (±1–15%) does not significantly affect the quality of the classification.*

***Key words**: saliva, lung cancer, classifier, random forest, cross-validation.*
(***For citation**:* Gundyrev IA, Bel'skaya LV, Kosenok VK, Sarf EA. The use of Synthetic Images for Solving the Classification Problem by the Example of Lung Cancer Diagnosis. *Annals of the Russian Academy of Medical Sciences*. 2018;73 (2):96–104. doi: 10.15690/vramn946)

биопсии опухоли. В каждом случае проведена гистологическая верификация диагноза, после чего принималось решение о включении или не включении пациента в исследование.

Исходы исследования

Основной исход исследования

Основным исходом исследования являлось определение принадлежности каждого участника исследования к классу «Здоров/Болен (рак легкого)» с использованием классификатора, построенного на данных биохимического анализа слюны.

Дополнительные исходы исследования

Дополнительным исходом исследования являлось установление зависимости результата классификации от вариативности исходных биохимических данных.

Анализ в подгруппах

Основная группа включала пациентов с гистологически подтвержденным диагнозом рака легкого. Контрольная группа включала условно здоровых пациентов, у которых при проведении плановой диспансеризации не было выявлено патологии легких. Контрольная группа разделена на 2 равные части для построения классификатора.

Методы регистрации исходов

В качестве материала для биохимических исследований использовали слюну. Во всех образцах определяли концентрацию сиаловых кислот, содержание белка, мочевины, активность щелочной фосфатазы, каталазы, содержание диазосоединений, уровень среднемолекулярных пептидов. Активность щелочной фосфатазы (Е/л) определяли методом конечной точки по Бессею–Лоури–Броку [27], каталазы (мкат/л) — по Королюку и соавт. [28]. Концентрацию мочевины (ммоль/л) определяли фотометрически уреазно-салицилатным методом по Бертлоту, общего белка (г/л) — по реакции с пирогаллоловым красным [29], диазосоединений (мкмоль/л) — по реакции диазотирования в присутствии сульфаниловой кислоты [30], сиаловых кислот (ммоль/л) — по методу Гесса [31]. Уровень молекул средней массы определяли методом ультрафиолетовой спектрофотометрии при длинах волн 254 и 280 нм [32]. Результаты выражали в единицах, количественно равных показателям экстинкции. Дополнительно оценивали значение коэффициента распределения как отношение экстинкций при длинах волн 280 и 254 нм соответственно.

Перечисленные параметры использованы для построения классификатора.

Выбор метода классификации

Для задачи бинарной классификации используется метрика качества — площадь под ROC-кривой (Area Under Curve, Receiver Operator Characteristic, AUC-ROC) [33]. На множестве базовых образов по информативным параметрам построены различные классификаторы: линейный дискриминантный анализ, наивный байесовский классификатор, метод опорных векторов (SVM), градиентный бустинг (GBM), случайный лес (Random Forest), метод k-ближайших соседей (kNN) [33–36], проведена кросс-валидация (при помощи библиотеки *caret*) — разбиение выборки на 5, 10 частей с сохранением соотношения классов. Лучшие результаты по метрике AUC-ROC были получены на классификаторе «случайный лес», близкие результаты — у GBM и SVM.

Использование синтетических образов для классификации

Обучение классификатора проводится на множестве базовых образов, затем на множестве синтетических образов сравниваются точность, чувствительность, специфичность. Для оценки качества работы классификатора используется кросс-валидация. Для этого проводится разбиение выборки на 5 и 10 частей с сохранением соотношения классов. Близкие образы либо все входят в тренировочную выборку, либо в тестовую, т.е. недопустимо использовать близкие образы и для обучения, и для проверки классификатора. В противном случае классификатор переобучается, т.е. на новых данных демонстрирует качество серьезно хуже, чем на тестовых. Кроме того, используется схема исключения по одному (LOOCV): в этом случае тестовая выборка состоит из одного базового образа и всех его близких, остальные образы составляют тренировочную выборку, на которой обучается классификатор. Если при проверке близких образов в созданном классификаторе часть из них попадает в один класс, а часть в другой, то считаем, что исходный образ затруднительно классифицировать однозначно, т.е. на таком образе классификатор работает неуверенно. Другой подход состоит в том, чтобы усреднять результаты по близким образам и присваивать класс базовому образу согласно тому, как классифицировано большинство близких образов, или усреднять численные результаты голосования на множестве близких образов.

Этическая экспертиза

Исследования одобрены на заседании комитета по этике БУЗ Омской области «Клинический онкологический диспансер» от 21 июля 2016 г., протокол № 15.

Статистический анализ

Принципы расчета размера выборки

Размер выборки предварительно не рассчитывался.

Методы статистического анализа данных

Статистический анализ выполнен при помощи программ Statistica 10.0 (StatSoft, США) и пакета R (версия 3.2.3) непараметрическим методом. Описание выборки производили с помощью подсчета медианы (Me) и интерквартильного размаха в виде 25-го и 75-го процентилей [LQ; UQ]. Различия считали статистически значимыми при $p<0,5$.

Результаты

Объекты (участники) исследования

В исследование включены 200 пациентов Клинического онкологического диспансера и 500 практически здоровых людей, выбранных в качестве контрольной группы. Основная группа включала 200 больных раком легкого с различными гистологическими типами, формой роста и стадией заболевания (табл. 1). Контрольная группа включала 500 условно здоровых пациентов, у которых при проведении плановой диспансеризации не было выявлено патологии легких. С целью получения объективных данных было построено 2 классификатора (200 больных и по 250 здоровых).

Основные результаты исследования

На начальном этапе получены значения чувствительности, специфичности, общей точности для построенных классификаторов при условии разбиения выборки на

Таблица 1. Характеристика пациентов для построения модели

Характеристики группы	Основная группа *n*=200 (%)	Контрольная группа *n*=500 (%)
Возраст, лет	58,5±0,9	49,4±4,7
Пол: мужчины	200 (100)	500 (100)
Гистологический тип:		
• аденокарцинома	90 (45)	-
• плоскоклеточный рак	80 (40)	-
• смешанный (аденокарцинома + плоскоклеточный рак)	5 (2,5)	-
• нейроэндокринный рак	25 (12,5)	-
Форма роста:		
• центральный рак	72 (36)	-
• периферический рак	118 (59)	-
• медиастинальный рак	10 (5)	-
Стадия заболевания:		
• St1	51 (25,5)	-
• St2	31 (15,5)	-
• St3	67 (33,5)	-
• St4	51 (25,5)	-
Наличие/отсутствие метастазов:		
• M0	140 (70)	-
• M1	60 (30)	-

5 и 10 частей, приблизительно равных по размеру, с сохранением соотношения размеров классов согласно исходной выборке (табл. 2).

Показано, что для исходной выборки на модельном классификаторе получены сопоставимые значения диагностических характеристик, причем разбиение выборки на большее число частей не дает статистически достоверного изменения исследуемых параметров. В связи с этим для дальнейших расчетов использованы средние значения чувствительности, специфичности и общей точности по каждому классификатору.

Дополнительные результаты исследования

На следующем этапе проведено создание синтетических образов для каждого участника исследования (10 и 30 соответственно) с учетом возможного отклонения от базовых значений от 1 до 15% (табл. 3), после чего повторно проведена процедура определения диагностических характеристик на модельных классификаторах (табл. 4, 5). При этом получены значения общей и уверенной классификации. В случае уверенной классификации близкие образы, которые по результатам классификации попадают в разные классы, удаляются, тогда как в случае общей — учитываются.

Таблица 2. Базовые образы при кросс-валидации (*k*) равной 5, 10; %

Характеристика	*k*=5	*k*=10	Среднее
Классификатор 1			
Общая точность	77,78 [74,44; 78,89]	77,78 [73,33; 82,22]	77,78 [73,33; 80,00]
Чувствительность	65,00 [62,50; 67,50]	70,00 [60,00; 75,00]	67,50 [60,00; 70,00]
Специфичность	88,00 [80,00; 90,00]	88,00 [80,00; 92,00]	88,00 [80,00; 92,00]
Классификатор 2			
Общая точность	77,78 [77,78; 80,00]	78,89 [77,78; 82,22]	77,78 [77,78; 82,22]
Чувствительность	72,50 [67,50; 72,50]	72,50 [65,00; 80,00]	72,50 [65,00; 80,00]
Специфичность	86,00 [82,00; 88,00]	86,00 [80,00; 88,00]	86,00 [80,00; 88,00]

Таблица 3. Пример классификации при создании синтетических образов (n=10)

Номер образца	Настоящий класс	Предсказанный класс	Вероятность отнесения к соответствующему классу	
			ZDOR	ZNO
782-S1*	ZNO	ZNO	0,074	0,926
782-S2	ZNO	ZNO	0,064	0,936
782-S3	ZNO	ZNO	0,046	0,954
782-S4	ZNO	ZNO	0,034	0,966
782-S5	ZNO	ZNO	0,084	0,916
782-S6	ZNO	ZNO	0,088	0,912
782-S7	ZNO	ZNO	0,054	0,946
782-S8	ZNO	ZNO	0,032	0,968
782-S9	ZNO	ZNO	0,058	0,942
782-S10	ZNO	ZNO	0,104	0,896
4335-S1*	ZNO	ZNO	0,424	0,576
4335-S2	ZNO	ZNO	0,468	0,532
4335-S3	ZNO	**ZDOR**	0,652	0,348
4335-S4	ZNO	ZNO	0,378	0,622
4335-S5	ZNO	ZNO	0,472	0,528
4335-S6	ZNO	**ZDOR**	0,554	0,446
4335-S7	ZNO	ZNO	0,446	0,554
4335-S8	ZNO	**ZDOR**	0,542	0,458
4335-S9	ZNO	ZNO	0,428	0,572
4335-S10	ZNO	**ZDOR**	0,540	0,460

Примечание. * — базовый образ. S2–S10 — синтетические образы.

Таблица 4. Результаты классификации при n=10, %

Характеристика		±1–5%	±5–10%	±10–15%
Классификатор 1				
Общая точность	О	76,67 [73,78; 79,33]	74,89 [73,00; 77,33]	73,11 [72,44; 76,89]
	У	80,26 [77,63; 83,78]	80,28 [79,03; 85,29]	82,14 [79,31; 85,19]
Чувствительность	О	67,00 [60,50; 73,00]	65,50 [62,00; 69,50]	65,75 [63,00; 67,50]
	У	68,75 [63,16; 75,00]	69,23 [66,67; 78,95]	75,00 [70,83; 77,78]
Специфичность	О	83,20 [78,00; 89,60]	82,80 [80,40; 84,80]	81,60 [79,00; 83,60]
	У	90,91 [84,21; 95,00]	93,33 [83,78; 95,00]	91,67 [87,88; 93,89]
Классификатор 2				
Общая точность	О	79,11 [75,22; 80,44]	77,78 [75,78; 81,44]	77,78 [75,11; 79,67]
	У	82,35 [80,00; 85,33]	84,29 [81,08; 87,50]	85,48 [83,33; 86,71]
Чувствительность	О	72,00 [68,25; 77,00]	72,50 [67,00; 77,50]	70,50 [66,75; 76,00]
	У	76,47 [68,42; 78,95]	77,42 [70,00; 83,33]	76,54 [70,83; 82,14]
Специфичность	О	86,00 [80,80; 87,60]	84,00 [80,80; 87,00]	82,40 [80,20; 85,80]
	У	90,70 [86,96; 95,00]	91,18 [87,18; 94,44]	92,31 [88,57; 94,44]

Примечание. О — общая классификация, У — уверенная классификация.

Таблица 5. Результаты классификации при n=30, %

Характеристика		±1–5%	±5–10%	±10–15%
Классификатор 1				
Общая точность	О	77,33 [73,93; 78,78]	75,11 [73,44; 78,15]	72,85 [71,19; 75,78]
	У	81,69 [77,03; 83,33]	83,87 [78,26; 86,21]	83,33 [80,79; 87,50]
Чувствительность	О	67,00 [61,17; 75,17]	66,58 [62,33; 70,17]	64,67 [60,00; 67,67]
	У	68,57 [64,86; 80,00]	71,43 [68,97; 80,00]	72,42 [70,00; 76,92]
Специфичность	О	81,27 [78,80; 87,67]	81,33 [78,13; 85,07]	79,93 [76,53; 82,53]
	У	89,47 [82,61; 92,50]	90,00 [86,49; 96,67]	93,10 [88,43; 100,0]
Классификатор 2				
Общая точность	О	78,15 [74,41; 81,93]	78,15 [75,93; 80,07]	76,67 [75,11; 78,30]
	У	82,50 [80,00; 85,71]	86,15 [83,87; 87,88]	86,79 [84,38; 89,80]
Чувствительность	О	72,17 [66,83; 75,33]	71,83 [65,58; 78,83]	71,25 [66,00; 73,50]
	У	76,47 [68,42; 77,96]	80,00 [74,07; 82,14]	81,82 [72,73; 83,33]
Специфичность	О	84,67 [81,47; 87,26]	83,47 [80,93; 86,67]	81,60 [79,20; 86,13]
	У	90,00 [86,05; 94,74]	92,86 [89,47; 94,44]	92,31 [88,57; 96,97]

Примечание. О — общая классификация, У — уверенная классификация.

Установлено, что при увеличении отклонений синтетических образов от базового значения диагностические характеристики, полученные на обоих классификаторах, ухудшаются. Однако результат уверенной классификации, напротив, дает более высокие значения. В целом, увеличение числа исходных данных для построения классификатора до 4500 (n=10) позволяет получить максимальные значения чувствительности (77,42%), специфичности (93,33%) и общей точности (85,48%), что превосходит результаты, полученные на исходной выборке (см. табл. 4). Дальнейшее увеличение количества синтетических образов до 13 500 (n=30) уже не дает существенного роста исследуемых параметров: максимальные значения составили 81,82% для чувствительности, 93,10% для специфичности и 86,79% для общей точности (см. табл. 5).

Обсуждение

Резюме основного результата исследования

Определены базовые диагностические характеристики модельного классификатора (чувствительность — 72,5%; специфичность — 86,0%). При увеличении отклонений синтетических образов от базового значения диагностические характеристики при общей классификации ухудшаются, в то время как результат уверенной классификации дает более высокие значения (чувствительность — 81,8%, специфичность — 93,1%). Полученные результаты позволяют сделать вывод, что измерительная погрешность в диапазоне (±1–15%) не оказывает существенного влияния на качество классификации.

Обсуждение основного результата исследования

Задача систем медицинской диагностики сводится к определению заболеваний, которыми возможно болен пациент, на основе данных о симптомах заболевания или результатов диагностических обследований. В зависимости от типа медицинских данных можно выделить два основных подхода к диагностике:

- диагностика с использованием теории вероятностей и математической статистики, основанная на анализе объективной статистической информации;
- диагностика с использованием искусственного интеллекта на основе субъектной информации, т.е. знаний и опыта группы врачей [37].

В целом при использовании методов диагностики, основанных на теории вероятности и математической статистики, процедура диагностики представляет собой «черный ящик», поскольку ее результаты не понятны ни врачу, ни пациенту [38]. Наиболее часто вычисляется условная вероятность наличия заболевания при заданном наборе диагностических характеристик на основании априорной вероятности, полученной на основе экспериментальных данных. Однако интерпретация полученных данных зачастую требует анализа и корректировки.

Следует отметить, что получение первичных данных для построения классификаторов — это длительная процедура, поскольку пациенты с нужными диагнозами в количестве, необходимом для получения статистически достоверных результатов, могут набираться годами. Описанный подход позволяет искусственно увеличить количество базовых образов для построения классификатора, одновременно улучшив получаемые диагностические характеристики.

Для диагностики рака легкого были апробированы и показали свою неэффективность такие методы, как рентгенография грудной клетки и цитологическое исследование мокроты [39]. В настоящее время для скрининга рака легкого рекомендована низкодозовая компьютерная томография грудной клетки, однако ее применение ограничено возрастной группой 55–74 года и целевой аудиторией — заядлые курильщики или отказавшиеся от курения менее 15 лет назад. Большие надежды возлагаются на выявление ранних молекулярных маркеров рака легкого (РЭА, Cyfra 21-1, CA72-4 — для аденокарциномы; Cyfra 21-1, SCC, РЭА — для плоскоклеточного и крупноклеточного рака; ProGRP, НСЕ, РЭА — для мелкоклеточного рака) [40]. Однако применение молекулярных маркеров зачастую ограничивается уточняющей диагностикой, оценкой эффективности лечения и прогноза течения опухолевого процесса, доклиническим выявлением развития рецидивов и только в ряде случаев используется для активного выявления рака. Поэтому для проведения диагностики рака легкого необходимо внедрение новых или расширение функциональных возможностей существующих методов.

В данной работе построен модельный классификатор для диагностики рака легкого, показывающий стабильные диагностические характеристики на уровне 70 и 88% для чувствительности и специфичности соответственно. Необходимо уточнить, что в задачи настоящего исследования не входила разработка метода диагностики, демонстрирующего более высокие диагностические характеристики по сравнению с существующими. Модельный классификатор был выбран в качестве примера, поскольку в настоящее время клинические рекомендации по диагностике рака легкого не включают ни одного метода, основанного на многомерной оценке результатов лабораторных тестов [41]. При этом для каждого пациента, результаты анализов которого использованы в качестве базовых образов для построения классификатора, определяющей является величина вероятности отнесения к определенному классу «Здоров/Болен». Отнесение к каждому классу происходит автоматически при значении вероятности более 0,50. Большинство синтетических образов, полученных при условии, что внесены небольшие отклонения в значения параметров, остается в том же классе, что и базовый образ, из которого они получены. Тем не менее есть базовые образы, которые уверенно (с вероятностью более 0,70) классифицируются ошибочно как на исходном классификаторе, так и при использовании синтетических образов. Однако базовые образы, на которых классификатор дает результат принадлежности к классу в диапазоне 0,45–0,55, затруднительно классифицировать однозначно, поскольку при малом отклонении параметров «близкий образ» тем же классификатором может быть отнесен к другому классу. Поэтому необходимо введение третьего класса в классификатор, так называемой серой зоны (0,4–0,6), т.к. вероятность постановки ошибочного диагноза в данной области существенно повышается [40]. Такая практика является распространенной в клинической лабораторной диагностике. При попадании результата классификации в «серую зону» пациенту может быть предложено провести повторную диагностику через определенный интервал времени.

Ограничения исследования

Ограничения исследования связаны с использованием для верификации работы модельного классификатора только метода кросс-валидации, тогда как более показательным было бы проведение слепого тестирования на независимой выборке пациентов, что планируется сделать на следующем этапе. Дальнейшие исследования предполагается провести при увеличении числа базовых обучающих образов, добавив в модельный классификатор новые признаки, а также их преобразовании и комбинации для построения классификатора, дающего более высокие диагностические характеристики.

Заключение

Используемые в медицинской диагностике значения диагностических характеристик представляют собой численные значения, четко привязанные к выборке, на базе которой они рассчитаны. Проверить стабильность полученных диагностических характеристик можно, получив новые данные (например, с использованием слепого тестирования на независимой группе), что зачастую сопряжено со значительным объемом лабораторных исследований, или с помощью методов математического моделирования (например, методом кросс-валидации). Предложенный в данной работе способ создания синтетических базовых образов не только позволяет наименее затратным способом увеличить выборку, но и получить на базе классификатора более высокие диагностические характеристики.

В целом, математические способы диагностики и прогнозирования позволяют значительно сократить затраты средств и времени на диагностику заболевания, увеличить точность диагностики и помогают врачу оперативно принимать решения. Именно поэтому задача разработки и внедрения этих методов в практическую деятельность врачей является актуальной.

Источник финансирования

Исследования выполнены при финансовой поддержке ООО «ХимСервис».

Конфликт интересов

Авторы данной статьи подтвердили отсутствие конфликта интересов, о котором необходимо сообщить.

Участие авторов

Гундырев И.А. — разработка концепции и дизайна исследования, разработка алгоритма, статистическая обработка данных; Бельская Л.В. — разработка концепции и дизайна исследования, выполнение лабораторных исследований, обработка полученных результатов, подготовка текста; Косенок В.К. — сбор и анализ клинических данных, редактирование; Сарф Е.А. — сбор материала, выполнение лабораторных исследований.

ЛИТЕРАТУРА

1. Карякина О.Е., Добродеева Л.К., Мартынова Н.А., и др. Применение математических моделей в клинической практике // *Экология человека*. — 2012. — №7 — С. 55–64. [Karyakina OE, Dobrodeeva LK, Martynova NA, et al. Use of mathematical models in clinical practice. *Ecology human*. 2012;(7):55–64. (In Russ).]

2. Халафян А.А. *Современные статистические методы медицинских исследований*. — М.: Изд-во ЛКИ; 2008. — 320 с. [Khalafyan AA. *Sovremennye statisticheskie metody meditsinskikh issledovanii*. Moscow: Izd-vo LKI; 2008. 320 p. (In Russ).]
3. Омирова Н.И., Палей М.Н., Евсюкова Е.В., Тишков А.В. Композиция деревьев решений для распознавания степени тяжести хронической обструктивной болезни легких // *Информационно-управляющие системы*. — 2014. — №5 — С. 115–118. [Omirova NI, Paley MN, Evsyukova EV, Tishkov AV. Composition of decision trees for severity of chronic obstructive pulmonary desease recognition. *Informatsionno-upravlyayushchie sistemy*. 2014;(5):115–118. (In Russ).]
4. Liang L, Cai F, Cherkassky V. Predictive learning with structured (grouped) data. *Neural Netw*. 2009;22(5-6):766–773. doi: 10.1016/j.neunet.2009.06.030.
5. Самаха Б.А., Шевякин В.Н., Разумова К.В., Кореневская С.Н. Использование интерактивных методов классификации для решения задач медицинского прогнозирования // *Фундаментальные исследования*. — 2014. — №1 — С. 33–37. [Samaha BA, Shevyakin VN, Razumova KV, Korenevskaya SN. Using of interactive classification methods for solving problems of medical prediction. *Fundamental'nye issledovaniya*. 2014;(1):33–37. (In Russ).]
6. Смагин С.В. Комплекс программ для индуктивного формирования баз медицинских знаний // *Программные продукты и системы*. — 2014. — №4 — С. 108–113. [Smagin SV. Kompleks programm dlya induktivnogo formirovaniya baz meditsinskikh znanii. *Programmnye produkty i sistemy*. 2014;(4):108–113. (In Russ).]
7. Sotiras A, Gaonkar B, Eavani H, et al. *Machine learning as a means toward precision diagnostics and prognostics*. In: Wu G, Shen D, Sabuncu M, editors. *Machine learning and medical imaging*. The Elsevier and MICCAI Society Book Series. Elsevier; 2016. pp. 299–334. doi: 10.1016/b978-0-12-804076-8.00010-4.
8. Chen H, Tan C, Lin Z, Wu T. The diagnostics of diabetes mellitus based on ensemble modeling and hair/urine element level analysis. *Comput Biol Med*. 2014;50:70–75. doi: 10.1016/j.compbiomed.2014.04.012.
9. Mohebian MR, Marateb HR, Mansourian M, et al. A Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR) using optimized ensemble learning. *Comput Struct Biotechnol J*. 2017;15:75–85. doi: 10.1016/j.csbj.2016.11.004.
10. Ле Н.В., Камаев В.А., Панченко Д.П., Трушкина О.А. Обзор подходов к проектированию медицинской системы дифференциальной диагностики // *Известия Волгоградского государственного технического университета*. — 2014. — Т.20. — №6 — С. 50–58. [Le NV, Kamaev VA, Panchenko DP, Trushkina OA. A review of the approaches to designing medical expert system on differential diagnosis. *Izvestiya Volgogradskogo gosudarstvennogo tekhnicheskogo universiteta*. 2014;20(6):50–58. (In Russ).]
11. Гундырев И.А., Бельская Л.В. *Использование синтетических образов для задачи медицинской диагностики рака легкого*. В кн.: *Материалы X международной научной конференции* / Под общей ред. В.П. Колосова. — Самара; 2016. — С. 8–11. [Gundyrev IA, Bel'skaya LV. *Ispol'zovanie sinteticheskikh obrazov dlya zadachi meditsinskoi diagnostiki raka legkogo*. In: (Conference proceedings) *Materialy X mezhdunarodnoy nauchnoy konferentsii*. Ed by V.P. Kolosov. Samara; 2016. pp. 8–11. (In Russ).]
12. Wu Yo, Wu Yi, Wang J, et al. An optimal tumor marker group-coupled artificial neural network for diagnosis of lung cancer. *Expert Syst Appl*. 2011;38(9):11329–11334. doi: 10.1016/j.eswa.2011.02.183.
13. Хайленко В.А., Давыдов М.И., Новиков А.М., Сперанский Д.Л. Клиническое значение определения сиаловых кислот у больных раком легкого // *Вестник онкологического научного центра Российской академии медицинских наук*. — 1991. — Т.2. — №1 — С. 25–27. [Khailenko VA, Davydov MI, Novikov AM, Speransky DL. Clinical value of sialic acids in lung cancer patients. *Herald of N.N. Blokhin Cancer Research Center RAMS*. 1991;2(1):25–27. (In Russ).]
14. Lemjabbar-Alaoui H, McKinney A, Yang Y-W, et al. Glycosylation alterations in lung and brain cancer. *Adv Cancer Res*. 2015;126:305–344. doi: 10.1016/bs.acr.2014.11.007.
15. Shamberger RJ. Serum sialic acid in normals and in cancer patients. *J Clin Chem Clin Biochem*. 1984;22(10):647–651.
16. Tran TT, Nguyen TMP, Nguyen BN, Phan VC. Changes of Serum Glycoproteins in Lung Cancer Patients. *J Proteomics Bioinform*. 2008;1(1):11–16. doi: 10.4172/jpb.1000004.
17. Сперанский В.В., Алехин Е.К., Петрова И.В., Алехин В.Е. О роли гистамина и антигистаминных препаратов в онкогенезе // *Медицинский вестник Башкортостана*. — 2010. — Т.5. — №4 — С. 151–156. [Speransky VV, Alekhin EK, Petrova IV, Alekhin VE. The role of histamine and antihistamine drugs in oncogenesis. *Bashkortostan medical journal*. 2010;5(4):151–156. (In Russ).]
18. Флеминг М.В., Климов В.В., Чердынцева Н.В. О взаимовлиянии аллергических реакций и злокачественных процессов (современное состояние проблемы) // *Сибирский онкологический журнал*. — 2005. — №1 — С. 96–101. [Fleming MV, Klimov VV, Cherdyntseva NV. O vzaimovliyanii allergicheskikh reaktsii i zlokachestvennykh protsessov (sovremennoe sostoyanie problemy). *Sibirskii onkologicheskii zhurnal*. 2005;(1):96–101. (In Russ).]
19. Keskinege A, Elgun S, Yilmaz E. Possible implications of arginase and diamine oxidase in prostatic carcinoma. *Cancer Detect Prev*. 2001;25(1):76–79.
20. Манина И.В., Перетолчина Н.М., Сапрыкина Н.С., и др. Перспективы применения антагониста Н2-гистаминовых рецепторов (циметидина) в качестве адъюванта биотерапии меланомы // *Иммунопатология, аллергология, инфектология*. — 2010. — №4 — С. 42–51. [Manina IV, Peretolchina NM, Saprykina NS, et al. Prospects of using antagonist histamine 42-receptor (cimetidinum) as adjuvant for melanoma biotherapy treatment. *Immunopatologiya, allergologiya, infektologiya*. 2010;(4):42–51. (In Russ).]
21. Lattermann R, Geisser W, Georgieff M, et al. Integrated analysis of glucose, lipid, and urea metabolism in patients with bladder cancer. Impact of tumor stage. *Nutrition*. 2003;19(7–8):589–592. doi: 10.1016/S0899-9007(03)00055-8.
22. Liu J, Duan Y. Saliva: a potential media for disease diagnostics and monitoring. *Oral Oncol*. 2012;48(7):569–577. doi: 10.1016/j.oraloncology.2012.01.021.
23. Malathi M, Shrinivas BR. Relevance of serum alkaline phosphatase as a diagnostic aid in lung pathology. *Indian J Physiol Pharmacol*. 2001;45(1):119–121.
24. Soini Y, Kaarteenaho-Wiik R, Paakko P, Kinnula V. Expression of antioxidant enzymes in bronchial metaplastic and dysplastic epithelium. *Lung Cancer*. 2003;39(1):15–22. doi: 10.1016/S0169-5002(02)00392-6.
25. Dayem AA, Choi HY, Kim JH, Cho SG. Role of oxidative stress in stem, cancers and cancer stem cells. *Cancers (Basel)*. 2010;2(2):859–884. doi: 10.3390/cancers2020859.
26. Панкова О.В., Перельмутер В.М., Савенкова О.В. Характеристика экспрессии маркеров пролиферации и регуляции апоптоза в зависимости от характера дисрегенераторных изменений в эпителии бронхов при плоскоклеточном раке легкого // *Сибирский онкологический журнал*. — 2010. — №5 — С. 36–41. [Pankova OV, Perelmuter VM, Savenkova OV. Characteristics of prolifeartion marker expression and apoptosis regulation depending on the character of disregenerator changes in bronchial epithelium of patients with squamous cell lung cancer. *Sibirskii onkologicheskii zhurnal*. 2010;(5):36–41. (In Russ).]
27. *Клиническая биохимия. Сборник инструкций*. — Новосибирск; 2011. — 132 с. [*Klinicheskaya biokhimiya. Sbornik instruktsii*. Novosibirsk; 2011. 132 p. (In Russ).]
28. Королюк М.А., Иванова Л.И., Майорова И.Г., Токарев В.Е. Метод определения активности каталазы // *Лабораторное*

дело. — 1988. — №1 — С. 16–19. [Korolyuk MA, Ivanova LI, Mayorova IG, Tokarev VE. Metod opredeleniya aktivnosti katalazy. *Lab Delo*. 1988;(1):16–19. (In Russ).]

29. Островский О.В., Храмов В.А., Попова Т.А. *Биохимия полости рта*. — Волгоград: Изд-во ВолГМУ; 2010. — 184 с. [Ostrovsky OV, Khramov VA, Popova TA. *Biokhimiya polosti rta*. Volgograd: Izd-vo VolGMU; 2010. 184 p. (In Russ).]
30. Храмов В.А., Пригода Е.В. Уровень аминоазота и имидазольных соединений в ротовой жидкости человека // *Стоматология*. — 2002. — Т.81. — №6 — С. 10–11. [Khramov VA, Prigoda EV. Uroven' aminoazota i imidazol'nykh soedineniy v rotovoy zhidkosti cheloveka. Stomatologiya. 2002;81(6):10–11. (In Russ).]
31. Романенко Е.Г., Руденко А.И. Методика определения сиаловой кислоты в слюне // *Світ медицини та біології*. — 2013. — Т.9. — №1 — С. 139–142. [Romanenko EG, Rudenko AI. Metodika opredeleniya sialovoi kisloty v slyune. *Svit meditsini ta biologii*. 2013;9(1):139–142. (In Russ).]
32. Гаврилов В.Б., Бидула М.М., Фурманчук Д.А., и др. Оценка интоксикации организма по нарушению баланса между накоплением и связыванием токсинов в плазме // *Клиническая лабораторная диагностика*. — 1999. — №2 — С. 13–17. [Gavrilov VB, Bidula MM, Furmanchuk DA, et al. Otsenka intoksikatsii organizma po narusheniyu balansa mezhdu nakopleniem i svyazyvaniem toksinov v plazme. *Klin Lab Diagn*. 1999;(2):13–17. (In Russ).]
33. Флах П. *Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных*. — М.: ДМК Пресс; 2015. — 399 с. [Flakh P. *Mashinnoe obuchenie. Nauka i iskusstvo postroeniya algoritmov, kotorye izvlekayut znaniya iz dannykh*. Moscow: DMK Press; 2015. 399 p. (In Russ).]
34. Мастицкий С.Э., Шитиков В.К. Статистический анализ и визуализация данных с помощью R [интернет]. 2014. [Mastitskiy SE, Shitikov VK. Statisticheskiy analiz i vizualizatsiya dannykh s pomoshch'yu R [Internet]. 2014. (In Russ).] Доступно по: http://r-analytics.blogspot.com. Ссылка активна на 12.03.2018.
35. Шитиков В.К., Мастицкий С.Э. Классификация, регрессия и другие алгоритмы Data Mining с использованием R [интернет]. 2017. [Shitikov VK, Mastitskiy SE. Klassifikatsiya, regressiya i drugie algoritmy Data Mining s ispol'zovaniem R [Internet]. 2017. (In Russ).] Доступно по: https://github.com/ranalytics/data-mining. Ссылка активна на 12.03.2018.
36. Джеймс Г., Уиттон Д., Хасти Т., Тибширани Р. *Введение в статистическое обучение с примерами на языке R*. Пер. с англ. С.Э. Мастицкого. — М.: ДМК Пресс; 2016. — 460 с. [James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning - with applications in R*. Transl. from English. Moscow: DMK Press; 2016. 460 p. (In Russ).]
37. Ле Н.В. Интеллектуальная медицинская система дифференциальной диагностики на основе экспертных систем // *Вестник Саратовского государственного технического университета*. — 2014. — Т.2. — №1 — С. 167–179. [Le NV. An intelligent medical differential diagnosis system based on expert systems. *Vestnik Saratovskogo gosudarstvennogo tekhnicheskogo universiteta*. 2014;2(1):167–179. (In Russ).]
38. Поворознюк А.И. Система поддержки принятия решения в медицине на основе синтеза структурированных моделей объектов диагностики // *Научные ведомости Белгородского государственного университета*. — 2009. — Т.12. — №15–1 — С. 170–176. [Povoroznyuk AI. Sistema podderzhki prinyatiya resheniya v meditsine na osnove sinteza strukturirovannykh modelei ob"ektov diagnostiki. *Nauchnye vedomosti Belgorodskogo gosudarstvennogo universiteta*. 2009;12(15–1):170–176. (In Russ).]
39. Давыдов М.И., Заридзе Д.Г. Скрининг злокачественных опухолей // *Вестник российского онкологического научного центра им. Н.Н. Блохина Российской академии медицинских наук*. — 2014. — Т.25. — №3–4 — С. 5–16. [Davydov MI, Zaridze DG. Skrining zlokachestvennykh opukholei. *Herald of N.N. Blokhin Cancer Research Center RAMS*. 2014;25(3–4):5–16. (In Russ).]
40. Сергеева Н.С., Маршутина Н.В., Солохина М.П., и др. Современные представления о серологических опухолеассоциированных маркерах и их месте в онкологии // *Успехи молекулярной онкологии*. — 2014. — №1 — С. 69–84. [Sergeeva NS, Marshutina NV, Solokhina MP, et al. Modern conceptions of serological tumor markers and their role in oncology. *Advances in molecular oncology*. 2014;(1):69–80. (In Russ).]
41. Федеральные клинические рекомендации по диагностике и лечению больных раком легкого. [Federal'nye klinicheskie rekomendatsii po diagnostike i lecheniyu bol'nykh rakom legkogo. (In Russ).] Доступно по: http://www.volgmed.ru/uploads/files/2014-11/34115-federalnye_klinicheskie_rekomendacii_po_diagnostike_i_lecheniyu_bolnyh_rakom_legkogo_2013_http_oncology-association_ru.pdf. Ссылка активна на 12.03.2018.

КОНТАКТНАЯ ИНФОРМАЦИЯ

Гундырев Иван Анатольевич, кандидат физико-математических наук, аналитик ООО «Три-Софт»
Адрес: 644099, Омск, ул. Гагарина, д. 14, оф. 702, **e-mail**: ivangundyrev@yandex.ru, **SPIN-код**: 3338-4901,
ORCID: http://orcid.org/0000-0002-9845-0039

Бельская Людмила Владимировна, кандидат химических наук, директор по науке ООО «ХимСервис», доцент кафедры химической технологии и биотехнологии Омского государственного технического университета
Адрес: 644050, Омск, Проспект Мира, д. 11, **e-mail**: ludab2005@mail.ru, **SPIN-код**: 4189-7899,
ORCID: http://orcid.org/0000-0002-6147-4854

Косенок Виктор Константинович, доктор медицинских наук, профессор, академик РАМТН, заведующий кафедрой онкологии с курсом лучевой терапии Омского государственного медицинского университета
Адрес: 644013, Омск, ул. Завертяева, д. 9, корп. 1, **тел**.: +7 (3812) 60-17-46, **e-mail**: vic.kos_senok@mail.ru,
SPIN-код: 4578-1551, **ORCID**: http://orcid.org/0000-0002-2072-2460

Сарф Елена Александровна, заведующая лабораторией ООО «ХимСервис»
Адрес: 644070, Омск, ул. А. Нейбута, д. 91а, **e-mail**: nemcha@mail.ru, **SPIN-код**: 9161-0264,
ORCID: http://orcid.org/0000-0003-4918-6937