(19) Евразийское патентное ведомство

(21) 201900375 (13) A1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОЙ ЗАЯВКЕ

(43) Дата публикации заявки
2020.05.29

(22) Дата подачи заявки
2019.08.16

(51) Int. Cl. ***G16H 50/00*** (2006.01)
G01N 33/574 (2006.01)

(54) СПОСОБ И СИСТЕМА ДЛЯ СКРИНИНГОВОГО ОПРЕДЕЛЕНИЯ ВЕРОЯТНОСТИ НАЛИЧИЯ РАКА ЛЕГКОГО

(31) **2018140406**
(32) **2018.11.15**
(33) **RU**

(71) Заявитель:
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ ПЕРВЫЙ МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ МЕДИЦИНСКИЙ УНИВЕРСИТЕТ ИМЕНИ И.М. СЕЧЕНОВА МИНИСТЕРСТВА ЗДРАВООХРАНЕНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ (СЕЧЕНОВСКИЙ УНИВЕРСИТЕТ) (ФГАОУ ВО ПЕРВЫЙ МГМУ ИМ. И.М. СЕЧЕНОВА МИНЗДРАВА РОССИИ (СЕЧЕНОВСКИЙ УНИВЕРСИТЕТ)) (RU)

(72) Изобретатель:
Глыбочко Петр Витальевич, Свистунов Андрей Алексеевич, Фомин Виктор Викторович, Копылов Филипп Юрьевич, Секачева Марина Игоревна, Паршин Владимир Дмитриевич, Гитель Евгений Павлович, Рагимов Алигейдар Алекперович, Поддубская Елена Владимировна (RU)

(74) Представитель:
Куприянова О.И. (RU)

(57) Изобретение относится к области медицины, а именно онкологии, и может быть использовано для скринингового определения вероятности наличия рака легкого или выявления данного онкологического заболевания на ранней стадии. Скрининговое определение вероятности наличия рака легкого основано на измерении уровня биомаркеров в образце биологической жидкости, полученном у субъекта: HE4, ApoA2, CYFRA.21.1, Ddimer, ApoA1, TTR, B2M, CA125, hsCRP, CEA, sVCAM.1, CA15.3, и информации о поле пациента, с последующей обработкой совокупности полученных данных с использованием по меньшей мере одной классификационной модели, обученной для определения вероятности наличия рака легкого.

201900375 A1

201900375 A1

МПК: G16H50/20, G16H50/30

СПОСОБ И СИСТЕМА ДЛЯ СКРИНИНГОВОГО ОПРЕДЕЛЕНИЯ ВЕРОЯТНОСТИ НАЛИЧИЯ РАКА ЛЕГКОГО

Область техники, к которой относится изобретение

Изобретение относится к области медицины, а именно онкологии, и может быть использовано для скринингового определения вероятности наличия рака легкого, в т.ч. немелкоклеточного, или выявления данного онкологического заболевания на ранней стадии.

Уровень техники

Злокачественные опухоли представляют собой одну из самых значимых проблем здравоохранения не только в России, но и во всем мире.

Онкологические заболевания являются второй по частоте причиной смерти в России. Средний показатель заболеваемости злокачественными новообразованиями в 2016г. составил 408,6 чел на 100000 населения. Средний показатель смертности – 201,6 чел на 100 000 населения. Абсолютное число умерших – 295 729 чел. Онкологическая заболеваемость растет во всем мире. За последние 10 лет она увеличилась более, чем на 20%.

Немелкоклеточный рак легкого стоит на первом месте по распространенности среди мужского населения индустриальных стран (Claudia Allemani, Hannah K Weir, Helena Carreira et al. Global surveillance of cancer survival 1995- 2009: analysis of individual data for 25 676 887 patients from 279 population-based registries in 67 countries (CONCORD-2). Lancet 2014;(November 26). doi:http://dx.doi.org/10.1016/S0140- 6736(14)62038-9). Статистика развитых стран мира свидетельствует о неуклонном росте впервые выявленных случаев рака легкого по сравнению со злокачественными опухолями любой другой локализации.

Рак легкого в России также занимает лидирующие позиции в структуре онкологической заболеваемости и смертности. Ежегодно в России заболевают раком легкого свыше 63000 человек, в том числе свыше 53000 мужчин. Более 20000 пациентов, или 34,2%, на момент постановки диагноза имеют распространенные стадии опухолевого процесса, при которых результаты лечения остаются неудовлетворительными. Анализ неудач хирургического лечения показал, что наиболее частой причиной смерти оперированных больных являются гематогенные метастазы (60-70%) и локо-регионарные рецидивы (30-40%).

Таким образом, разработка новых доступных скрининговых способов ранней

диагностики рака легкого является очень актуальной задачей.

Обычной стратегией скрининга является ежегодная рентгенография грудной клетки, особенно у курильщиков. Однако, в крупном клиническом исследовании PLCO (Prostate, Lung, Colorectal and Ovarian Cancer screening) показано, что проведение такого скрининга не влияет на смертность от рака легких в популяции обследуемых (Oken et al., Screening by chest radiograph and lung cancer mortality: the Prostate, Lung, Colorectal, and Ovarian (PLCO) randomized trial. JAMA. 2011 Nov 2;306(17):1865-73). Это утверждение безусловно относится и к используемому в России методу рентгеновской диагностики – флюорографии. Исследование выполняется в одной проекции и, несомненно, еще менее информативно по сравнению с рентгенографией грудной клетки.

В настоящее время наиболее эффективным методом скрининговой диагностики в мире является низкодозная спиральная компьютерная томография (НДСКТ). При проведении крупного клинического исследования NLST (National Lung Screening Trial) было установлено, что ежегодная НДСКТ приводит к снижению на 20% смертности от рака легкого по сравнению с ежегодной рентгенографией грудной клетки (National Lung Screening Trial Research Team et al., Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011 Aug 4;365(5):395-409). Методика рекомендована для скрининга рака легких в США. Американская ассоциация торакальной хирургии рекомендует ежегодный скрининг в возрастной группе от 50 до 79 лет, у пациентов со стажем курения 20 лет и дополнительными сопутствующими заболеваниями, которые повышают общий риск развития рака на 5% в течение ближайших 5 лет.

Однако качественный скрининг методом КТ возможен только при наличии высококвалифицированных специалистов и современных аппаратов, которые есть только в крупных медицинских учреждениях. Негативным фактом является и то, что проведение повторных исследований связано с риском дополнительного облучения.

В качестве альтернативы вышеизложенным инструментальным методам визуализации могут выступать методы диагностики, основанные на определении биохимических маркеров в биологических тканях и жидкостях пациента, например, цельной крови, сыворотке или плазме. В качестве таких маркеров, например, могут быть использованы различные антигены, протеины и метаболиты, секретируемые злокачественными клетками или образующиеся в процессе их гибели. Так, в настоящее время для диагностики рака легкого наиболее широко используется определение CYFRA 21-1 (фрагмент цитокератина 19) и CEA (раковый эмбриональный антиген) в плазме крови, известны и другие биомаркеры (Zamay et al., Current and Prospective Protein

Biomarkers of Lung Cancer. Cancers 2017, 9, 155). Стоит отметить, что диагностика онкологических заболеваний на основе измерений единичных биомаркеров не является достаточно достоверной ввиду их невысокой чувствительности. Так, например, чувствительность и специфичность CYFRA 21-1 в диагностике рака легкого составляет 43% и 89%, CEA – 69% и 68% соответственно (Zamay et al., Current and Prospective Protein Biomarkers of Lung Cancer. Cancers 2017, 9, 155). Использование мультиплексных диагностических методов, подразумевающих оценку риска наличия заболевания на основе измерений нескольких биомаркеров, позволяет преодолеть данную проблему и достичь более достоверных результатов.

Так, например, из KR-10-2016-0113444 (прототип) известно определение наличия рака легкого по измеренным в сыворотке крови маркерам следующих белков: HE4, RANTES, sVCAM-1, LRG1, CEA, CYFRA 21-1, ApoA2, ApoA1, TTR, B2M, CA125, CA19-9, hsCRP. При этом риск наличия заболевания оценивается по методу логит-регресии на основании совокупности измерений вышеизложенных биомаркеров.

Несмотря на возможность использования в методике комплекса маркеров, повышающих ее диагностическую ценность при оценке риска развития рака легкого, существует необходимость в адаптации методики для различных групп обследуемых. В литературе отмечены межрассовые различия в молекулярных механизмах рака легкого, что ставит под сомнение целесообразность использования единого набора биомаркеров для разных рас. Так, было показано, что частота встречаемости мутаций рецептора эпидермального фактора роста выше в азиатской популяции по сравнению с европеоидной, в то время как частота встречаемости KRAS мутаций – ниже (M.B. Schabath, D. Chress, T. Munoz-Antonia. Racial and Ethnic Differences in the Epidemiology and Genomics of Lung Cancer. Cancer Control. 2016 Oct;23(4):338-346). Подобные межрасовые различия могут быть вызваны как факторами окружающей среды (уровень загрязнения воздуха), поведенческими особенностями (специфика питания, распространенность курения), так и генетической предрасположенностью (W. Zhou and D. C. Christiani. East meets West: ethnic differences in epidemiology and clinical behaviors of lung cancer between East Asians and Caucasians. Chin J Cancer. 2011 May; 30(5): 287–292).

Заявляемое изобретение основано на исследовании нового комплекса маркеров, позволяющего повысить точность и достоверность определения наличия заболевания при скрининге рака легкого у конкретного пациента европеоидной популяции, формирование на этой основе той или иной группы риска и выявление тех пациентов, которые нуждаются в углубленном дорогостоящем обследовании для обнаружения ранней стадии

рака легкого.

Раскрытие изобретения

Технической проблемой, решаемой настоящим изобретением, является создание более точного способа определения вероятности наличия рака легкого в европеоидной популяции.

Достигаемым техническим результатом является повышение точности скринингового выявления наличия рака у конкретного пациента европеоидной популяции, причем уже на ранних стадиях его развития посредством выявления и учета оригинальной совокупности биомаркеров по итогам анализа фракции сыворотки или плазмы крови при ускорении диагностируемых состояний.

Технический результат достигается посредством реализации способа скринингового определения вероятности наличия рака легкого, включающего измерение уровня биомаркеров в образце биологической жидкости, полученном у субъекта: HE4, ApoA2, CYFRA.21.1, Ddimer, ApoA1, TTR, B2M, CA125, hsCRP, CEA, sVCAM.1, CA15.3, а также определение пола пациента, с последующей обработкой совокупности полученных значений биомаркеров с использованием, по меньшей мере, одной классификационной модели, обученной для определения высокой или низкой вероятности наличия рака легкого.

В качестве классификационных моделей используют метод «случайного леса» (random forest), и/или линейный дискриминантный анализ, и/или метод опорных векторов.

Обученную классификационную модель получают посредством реализации следующих шагов:

- формируют обучающую и тестовую выборку записей субъектов с измеренными значениями биомаркеров HE4, ApoA2, CYFRA.21.1, Ddimer, ApoA1, TTR, B2M, CA125, hsCRP, CEA, sVCAM.1, CA15.3), включающие записи о пациентах разного пола и возраста;

- обучают классификационную модель выявлению заданной патологии, используя записи обучающей и тестовой выборки;

- сохраняют связи и веса обученной классификационной модели, для последующего определения вероятности наличия рака легкого по итогам обработки измеренных данных биомаркеров субъекта.

При формировании обучающей и тестовой выборки, включают записи субъектов с выявленной патологией - наличие рака и отсутствие рака легкого.

Технический результат достигается посредством реализации **системы**

скринингового определения вероятности наличия рака легкого, включающей

- модуль ввода измеренных значений биомаркеров субъекта;

- модуль хранения данных, выполненный с возможностью хранения обучающей и тестовой выборки классификационной модели, связей и весов обученной классификационной модели, записей субъектов с измеренными значениями биомаркеров HE4, ApoA2, CYFRA.21.1, Ddimer, ApoA1, TTR, B2M, CA125, hsCRP, CEA, sVCAM.1, CA15.3, включающие записи о пациентах разного пола и возраста;

- модуль обученной классификационной модели, выполненный с возможностью построения и обучения, по меньшей мере, одной классификационной модели для определения наличия заданной патологии по упомянутым маркерам, взятым из модуля хранения данных;

- модуль диагностики, выполненный с возможностью обработки введенных значений биомаркеров субъекта с использованием, по меньшей мере, одной обученной классификационной модели;

- модуль вывода данных, выполненный с возможностью получения данных о высокой или низкой вероятности наличия рака легкого.

Точность заявляемого мультиплексного метода диагностики рака легкого обеспечивается за счет использования комплекса из 12 биомаркеров и информации о поле пациента, а также за счет использования нескольких классификационных моделей с последующим усреднением модельных результатов.

Краткое описание чертежей

Изобретение поясняется чертежами, где:

На фиг.1 А. представлена диаграмма рассеяния «возраст пациента - концентрация биомаркеров». Точки – индивидуальные измерения, линии – предсказания линейной регрессионной модели. На графиках приведены значения корреляционных коэффициентов, рассчитанных по методу Пирсона и Р-значения, рассчитанные по тесту Стьюдента; на фиг.1 Б. представлена диаграмма размаха для оценки значимости гендерных различий в концентрациях биомаркеров. На графиках приведены Р-значения, полученные при помощи критерия Стьюдента. Серым цветом показаны данные для женщин, черным – для мужчин;

На фиг.2 - ROC-кривые для оценки предсказательной способности отдельных биомаркеров (тип линий соответствует биомаркеру);

На фиг.3. - Примеры деревьев решений, полученных в результате обучения многофакторного классификационного алгоритма random forest на экспериментальных

данных по 12 биомаркерам;

На фиг. 4 - Визуализация результатов разделения пациентов на 2 класса (здоровые доноры и пациенты с раком легкого) при помощи линейного дискриминантного анализа по 12 биомаркерам;

На фиг.5 - Примеры 3-мерных проекций разделения объединенной популяции пациентов на 2 класса (здоровые доноры и пациенты с раком легкого) при помощи метода опорных векторов по 12 биомаркерам;

На фиг.6 - Доля классификаторов стратифицированная по AUROC в зависимости от количества включенных в них биомаркеров. Обучение проводилось при помощи А. Метода опорных векторов Б. Линейного дискриминантного анализа.

На фиг.7 - ROC-кривые для оценки предсказательной способности различных классификационных алгоритмов. А. Весь набор данных был использован как для обучения модели, так и для ее валидации; Б. 80% данных было использовано для обучения модели, 20% - для валидации.

На фиг.8 - Блок-схема системы, предназначенной для оценки вероятности наличия рака легкого на основе данных пациента.

На фиг.9 - Алгоритм оценки вероятности наличия рака легкого на основе данных пациента.

Осуществление изобретения

Исходная группа биомаркеров, используемая в диагностическом тесте на определение вероятности наличия рака легкого (РЛ) была получена с использованием многофакторной классификационной модели. Подобные методы позволяют находить комбинации биомаркеров, обладающих наибольшим диагностическим потенциалом. Математическая модель проходит обучение на экспериментальных измерениях заданного набора биомаркеров, полученных на смешанной выборке из здоровых добровольцев и пациентов с РЛ. Обученная модель может быть использована для оценки риска наличия заболевания у пациента на основе показателей его биомаркеров.

В рамках проведенной работы на этапе разработки диагностически значимого комплекса показателей были использованы данные измерений 16 биомаркеров (AFP, CEA, CA 19-9, CA 125, HE4, tPSA, CA 15-3, B2M, hsCRP, Ddimer, CYFRA 21-1, ApoA1, ApoA2, Apo B, TTR, sVCAM-1), полученные на выборке здоровых добровольцев европеоидной популяции (n=203, 104 женщины и 99 мужчин 36-80 лет, средний возраст 53 года) и пациентов с раком легкого (n=77, 25 женщин и 52 мужчин 36-80 лет, средний возраст 62 лет).

Статистическая обработка экспериментальных данных и разработка классификационных моделей проводилась в среде R {RDevelopmentCoreTeam (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0}.

Первым этапом являлся статистический анализ и визуализация данных. Для здоровых добровольцев было оценено влияние пола и возраста на показатели биомаркеров (Фиг.1).

По итогам проведения исследования на данном этапе был сделан вывод об отсутствии значимой корреляции между возрастом пациентов и показателями большинства биомаркеров, в то же время наблюдались значимые гендерные различия в показателях CEA, CA 19-9, CA 125, HE4, Ddimer, ApoA1 и TTR .

На следующем этапе проводилась оценка значимости различий в уровнях отдельных биомаркеров между здоровыми добровольцами и пациентами с раком легкого при помощи критерия Стьюдента после нормализации экспериментальных данных путем log-трансофрмации (Таблица 1).

Таблица 1. Сравнение показателей биомаркеров в плазме крови здоровых доноров и пациентов с раком легкого

Биомаркер	Здоровые доноры	Пациенты с раком легкого	P-значения[1]	P-значения[1] (прототип)
AFP, МЕ/мл	2.84003	2.857935	0.3	-
ApoA1, г/л	1.59803	1.718442	4e-04 ***	<2.2e -16 ***
ApoA2, г/л	0.2953	0.218312	1e-13 ***	<2.2e -16 ***
ApoB, г/л	1.026404	0.94039	0.008**	-
B2M, нг/мл	1477.291	1887.649	3e-09 ***	<2.2e -16 ***
CA125, МЕ/мл	10.68987	44.81299	4e-06 ***	2.2e -14 ***
CA15.3, МЕ/мл	15.12623	26.86701	0.004 **	-
CA19.9, МЕ/мл	6.597143	13.31935	0.08	0.2829
CEA, нг/мл	1.856232	18.98655	2e-05 ***	<2.2e -16 ***
CYFRA.21.1, нг/мл	1.374517	5.257714	2e-11 ***	<2.2e -16 ***
Ddimer, нг/мл	119.7537	462.6623	3e-13 ***	-
HE4, пмоль/л	51.42532	104.9732	8e-21 ***	<2.2e -16 ***
hsCRP, мг/л	1.770049	11.55325	5e-08 ***	<2.2e -16 ***
sVCAM.1, нг/мл	658.3713	815.4416	2e-04 ***	0.01928*
t.PSA, нг/мл	1.129717	1.735423	0.3	-
TTR, мг/дл	25.64039	18.96104	2e-09 ***	<2.2e -16 ***

[1] – сравнение по тесту Стьюдента после нормализации экспериментальных данных; * - P-значения<0.05, ** - P-значения<0.01, *** - P-значения<0.001.

На основании проведенного анализа был сделан вывод об отсутствии значимых различий в концентрациях AFP, CA19.9, t.PSA между здоровыми добровольцами и пациентами с раком легкого. Так же в рамках данного исследования отмечено значимое различие в концентрациях Ddimer, ApoB и CA15.3, не включенных в прототип, а также отмечено более значимое различие в sVCAM.1 по сравнению с исследованием-прототипом.

Для оценки диагностической ценности отдельных биомаркеров использовался метод логистических регрессий. В данных статистических моделях рассматривалась взаимосвязь между концентрацией биомаркера и вероятностью наличия заболевания (уравнение 1):

$$P(Y) = {}^{1}/_{(1 + e^{-(b_0 + b_1 \cdot X)})} \quad (1)$$

где P(Y) – вероятность наличия заболевания, b_0 и b_1 – коэффициенты, определяемые по экспериментальным данным, X – предиктор (концентрация биомаркера).

Предсказательная способность логистических моделей оценивалась при помощи ROC-анализа, предполагающего определение чувствительности, специфичности и точности метода относительно тестового или общего набора данных. Для этого значение пороговой вероятности, определяющей наличие заболевания, варьировалось в пределах от 0 до 1 с заданным шагом, для каждого шага рассчитывалась доля верно диагностированных случаев заболевания (чувствительность) (S_e), правильно определенных случаев отсутствия заболевания (специфичность) (S_p), а также общая доля правильно диагностированных случаев, как наличия, так и отсутствия заболевания (точность) (Acc), (уравнения 2-4):

$$S_e = {}^{TP}/_{(TP + FN)} \cdot 100\% \quad (2)$$

$$S_p = {}^{TN}/_{(TN + FP)} \cdot 100\% \quad (3)$$

$$Acc=(TN+TP)/(TN+FP+TP+FN)\cdot 100\% \quad (4)$$

где TP – верно классифицированный положительный результат (верно диагностированное заболевание), FP – ложноположительный результат (ошибочно диагностированное заболевание), TN – верно классифицированный отрицательный результат (верно диагностированное отсутствие заболевания), FN – ложноотрицательный результат (ошибочно диагностированное отсутствие заболевания).

Полученный набор значений чувствительности и специфичности использовался для построения ROC-кривой. В качестве интегрального показателя качества моделей использовалась площадь под ROC-кривой (AUROC): предикторы с максимальной

предиктивной способностью показывают наибольшие значения AUROC. Результаты ROC-анализа приведены на фиг. 2 и в таблице 2.

Таблица 2. Диагностическая ценность отдельных биомаркеров

Биомаркер	Специфичность, %	Чувствительность, %	Точность, %	AUROC
HE4	90	82	88	0.9
ApoA2	97	74	90	0.88
CYFRA.21.1	91	61	82	0.83
Ddimer	67	84	71	0.82
ApoA1	81	70	78	0.81
TTR	85	60	78	0.8
B2M	58	83	65	0.76
CA125	82	56	75	0.73
hsCRP	85	56	77	0.72
CEA	83	52	75	0.68
sVCAM.1	88	45	76	0.65
CA15.3	69	49	64	0.6
ApoB	53	66	56	0.6
CA19.9	71	47	64	0.59
AFP	48	69	54	0.56
t.PSA	86	40	70	0.56

На основе результатов статистического анализа данных и оценки предсказательной способности однофакторных логистических моделей были отобраны биомаркеры, которые впоследствии были включены в многофакторные классификационные модели. Критерием включения биомаркеров являлись pval<0.005 (Таблица 1) и AUROC≥0.6 (Таблица 2). Таким образом, для построения классификационных моделей были отобраны экспериментальные измерения 12 биомаркеров (**HE4, ApoA2, CYFRA.21.1, Ddimer, ApoA1, TTR, B2M, CA125, hsCRP, CEA, sVCAM.1, CA15.3**), так же использовалась информация о поле пациента.

Разработка многофакторных классификационных моделей являлась завершающим этапом исследования. Различные способы машинного обучения (random forest, линейный дискриминантный анализ, метод опорных векторов) были использованы в рамках текущей задачи. Оценка параметров моделей (обучение), производилась на объединённых данных, полученных на здоровых добровольцах и пациентах с раком легкого, и была направлена на минимизацию предсказательных ошибок алгоритма. Детальное описание использованных методов изложено в книге (Bishop CM, Pattern recognition and machine learning. Springer. 2006).

Метод «random forest» (RF) подразумевает создание совокупности кросс-валидированных решающих деревьев. Каждое из таких деревьев проходит обучение на

подвыборке данных, включающей информацию лишь по части биомаркеров и наблюдений, и валидируется на подвыборке, не использованной для его построения (бэггинг). На основании предсказаний каждого из построенных деревьев решений пациент причисляется к одной из групп (здоровые доноры или пациенты с раком легкого), финальное предсказание классификатора определяется большинством голосов построенных деревьев (см. фиг.3 А, Б).

Использование линейного дискриминантного анализа (LDA) предполагает поиск линейной комбинации биомаркеров - дискриминанты, обеспечивающей наилучшее разделение всей популяции обследуемых на здоровых добровольцев и пациентов с раком легкого. Линейная дискриминанта может быть рассчитана: $z(x) = \beta_1 x_1 + \cdots + \beta_n x_n$, где x_i– это концентрации i-го биомаркера, β_i – коэффициенты модели. Данная задача решается за счет нахождения оси, проекция на которую обеспечивает максимальное отношение общей дисперсии линейной комбинации биомаркеров выборки к сумме дисперсий линейной комбинации биомаркеров внутри классов (см.фиг.4).

Таблица 3. Значения линейных коэффициентов при дискриминанте (LDAcomponent 1)

Фактор	Коэффициент
SEX(M)	1.78E-01
CEA	-3.53E-03
CA125	7.98E-03
HE4	2.48E-02
CA15.3	-9.30E-03
B2M	3.23E-04
hsCRP	-1.58E-03
Ddimer	-3.00E-05
CYFRA.21.1	-5.24E-02
ApoA1	1.94E-02
ApoA2	-8.02E+00
TTR	-6.34E-02
sVCAM.1	9.86E-05

Использование метода опорных векторов (SVM) предполагает нахождение (n-1)-мерной гиперплоскости, разделяющей n-мерное пространство значений биомаркеров на два класса. Пусть имеется обучающая выборка $(x_1, y_1), \ldots (x_n, y_n), x_i \in R^n, y_i \in \{-1,1\}$, где x_i— это вектор значений биомаркеров, а y_i определяет принадлежность пациента к классу. Классифицирующая функция может быть определена как $F(x) = \mathrm{sign}(\langle w, x\rangle + b)$, где w — нормальный вектор к разделяющей гиперплоскости, b — вспомогательный параметр, а фунуция может принимать значения 1 или -1 в зависимости от класса объекта. Обучение алгоритма подразумевает поиск такой гиперплоскости, которая обеспечивает

наименьшую эмпирическую ошибку классификации и максимизирует расстояние между значениями биомаркеров пациентов, относящихся к разным классам (см. фиг.5):

На первом этапе построения многофакторных моделей проводилось изучение диагностической ценности различных комбинаций биомаркеров из приведенной выше группы. Для этого все возможные комбинации, включающие от 2 до 12 биомаркеров были использованы для построения классификационных моделей (4803 варианта). Для обучения использовались объединённые данные, полученные на здоровых добровольцах и пациентах с раком легкого, и методы линейного дискриминантного анализа и опорных векторов. Разработанные модели были ранжированы в соответствии с их предсказательным потенциалом, оцененным по показателю AUROC (фиг. 7, таблица 4).

Как видно из фиг. 6, набольшей предсказательной способностью обладают комплексные тесты, включающие 11-12 биомаркеров, в то время как для относительно небольшой доли классификаторов, включающих комбинации из 2-3 биомаркеров, показатель AUROC составляет более 80%.

Финальной фазой построения классификаторов являлась их валидация.

Объединённые данные, полученные на здоровых добровольцах и пациентах с раком легкого были случайным образом разделены на обучающую и тестовую выборки. Оценка параметров моделей (обучение), производилась на обучающей выборке и была направлена на минимизацию предсказательных ошибок алгоритма. Валидация обученных моделей заключалась в оценке их предсказательной способности на тестовой выборке. Предсказательная способность многофакторных классификационных моделей оценивалась при помощи ROC-анализа как это было сделано ранее для отдельных биомаркеров (фиг. 7, Таблица 4).

Таблица 4. Диагностическая ценность обученных многофакторных классификационных моделей

Метод	Чувствительность, %	Специфичность, %	Точность, %	Пороговая вероятность, %	AUROC
Весь набор данных был использован как для обучения модели, так и для ее валидации					
RF	100	100	100	50	1
LDA	69	97	89	50	0.93
SVM	79	98	92	50	0.99
80% данных было использовано для обучения модели, 20% - для валидации					

RF	79	100	93	50	0.99
LDA	58	95	83	50	0.94
SVM	74	95	88	50	0.98

Финальные классификационные модели представляют собой обученные алгоритмы, позволяющие предсказать вероятность наличия рака легкого на основании экспериментальных измерений биомаркеров пациентов с учетом гендерных различий.

Финальное решение - определение вероятности наличия рака легкого, рассчитывается как медиана значений вероятностей рака легкого, рассчитанных в 3 классификационных моделях (RF, LDA SVM), обученных на всей выборке пациентов (см., например, Kittler J, Hatef M, Duin RPW et al, On Combining Classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 20, NO. 3, MARCH 1998 226-39.)

Для реализации заявляемого способа было разработано программное обеспечение (ПО), позволяющее на основе данных конкретного пациента (пол и результаты измерения биомаркеров) рассчитывать вероятность наличия у него рака легкого. Блок-схема реализации изобретения представлена на фиг. 8.

Компьютерно-реализуемая система состоит из (1) интерфейса, включающего устройство ввода данных пациента (пол и результаты измерений биомаркеров) и вывода результатов расчета (вероятность наличия рака легкого); (2) блока памяти, содержащего обученные классификаторные модели и программные продукты, необходимые для работы с ними (R portable, Google Chrome Portable) и (3) программного модуля, с помощью которого реализуется программный код, необходимый для обмена данных между интерфейсом и блоком памяти. Для создания графического интерфейса был использован пакет shiny (Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2017). shiny: Web Application Framework for R. R package version 1.0.5. https://CRAN.R-project.org/package=shiny) созданный на базе среды R {RDevelopmentCoreTeam (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0}. Для работы с данным пакетом необходимо наличие программных продуктов R portable и Google Chrome portable, хранящихся в блоке памяти. Для работы с предложенными моделями необходимы следующие пакеты: (1) RandomForest (A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18-22); (2) MASS (Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0); (3) e1071 (David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2017).

e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-8. https://CRAN.R-project.org/package=e1071).

Алгоритм оценки вероятности наличия рака легкого на основе данных пациента представлен на фиг.9.

Данные пациента вводятся через интерфейс и подаются в качестве входных переменных в разработанные модели, в каждой из которых производится расчет вероятности наличия рака легкого. Далее по результатам модельных предсказанной рассчитывается среднее значение, которое выводится в окно вывода.

Диагностическая мультиплексная панель для оценки риска рака легкого включает биомаркеры, показавшие максимальный предсказательный потенциал в рамках проводимого исследования (рис. 2, таблица 2): HE4, ApoA2, CYFRA.21.1, Ddimer, ApoA1, TTR. Кроме того, в заявляемый комплекс включены дополнительные биомаркеры, обладающие меньшим предсказательный потенциалом, однако значимо различные между здоровыми добровольцами и пациентами с рака легкого (Таблица. 1): B2M, CA125, hsCRP, CEA, sVCAM.1 и CA15.3 в исследуемой популяции.

Ниже представлено описание одного из клинических примеров применения способа, подтверждающего возможность реализации изобретения с достижением технического результата.

Пример 1 Больной К., 54 лет.

Курит 35 лет.

В январе 2018 года в связи с жалобами на слабость и быструю утомляемость обратился в поликлинику по месту жительства.

Был осмотрен терапевтом. Рекомендован прием витаминов, общий анализ крови, в котором клинически значимых отклонений не было выявлено.

Пациенту было предложено принять участие в программе Онкопоиска.

Пациент обследован в рамках программы. Получены следующие результаты исследования сыворотки крови: AFP 2,4 МЕ/мл, CEA 2,1 нг/мл, CA 19-9 3,6 МЕ/мл, CA 125 9,7 МЕ/мл, HE4 110,2 пмоль/л, tPSA 0,65 нг/мл, CA 15-3 19,2 МЕ/мл, B2M 2154нг/мл, hsCRP <0,08 нг/мл, D-dimer 51,0 ,CYFRA 21-1 1,28 нг/мл, Apo A-1 1,38 г/л, Apo A2 0,289 г/л, Apo B 1,15 г/л, TTR (prealb) 25,0 мг/дл, sVCAM-1 812 нг/мл, Rantes 40784 пг/мл, VEGFR1 135 пг/мл.

При обработке полученных результатов заявляемым способом выявлена высокая вероятность рака легкого.

Выполнена РКТ с контрастированием. Выявлено образование нижней доли правого легкого 13х12мм, с неровными тяжистыми контурами, неоднородно накапливающее контрастный препарат. Лимфоузлы средостении не увеличены. Пациент госпитализирован для хирургического лечения. Выполнена видеоассистированная торакоскопия, резекция нижней доли правого лекгого, медиастинальная лимфодиссекция. Гистол.№ высокодифференцированная аденокарцинома легкого. В 5 удаленных л/узлах – без признаков метастатического роста.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. **Способ скринингового определения вероятности наличия рака легкого,** включающий измерение уровня биомаркеров в образце биологической жидкости, полученном у субъекта: HE4, ApoA2, CYFRA.21.1, Ddimer, ApoA1, TTR, B2M, CA125, hsCRP, CEA, sVCAM.1, CA15.3, а также определение пола пациента, с последующей обработкой совокупности полученных значений биомаркеров с использованием, по меньшей мере, одной классификационной модели, обученной для определения высокой или низкой вероятности наличия рака легкого.

2. Способ по п.1, характеризующийся тем, что в качестве классификационных моделей используют метод «случайного леса» (random forest), и/или линейный дискриминантный анализ, и/или метод опорных векторов.

3. Способ по п.1, характеризующийся тем, что обученную классификационную модель получают посредством реализации следующих шагов:

- формируют обучающую и тестовую выборку записей субъектов с измеренными значениями биомаркеров HE4, ApoA2, CYFRA.21.1, Ddimer, ApoA1, TTR, B2M, CA125, hsCRP, CEA, sVCAM.1, CA15.3), включающие записи о пациентах разного пола и возраста;

- обучают классификационную модель выявлению заданной патологии, используя записи обучающей и тестовой выборки;

- сохраняют связи и веса обученной классификационной модели, для последующего определения вероятности наличия рака легкого по итогам обработки измеренных данных биомаркеров субъекта.

4. Способ по п. 3, характеризующийся тем, что при формировании обучающей и тестовой выборок, включают записи субъектов с выявленной патологией - наличие рака и отсутствие рака легкого.

5. **Система скринингового определения вероятности наличия рака легкого,** включающая

- модуль ввода измеренных значений биомаркеров субъекта HE4, ApoA2, CYFRA.21.1, Ddimer, ApoA1, TTR, B2M, CA125, hsCRP, CEA, sVCAM.1, CA15.3;

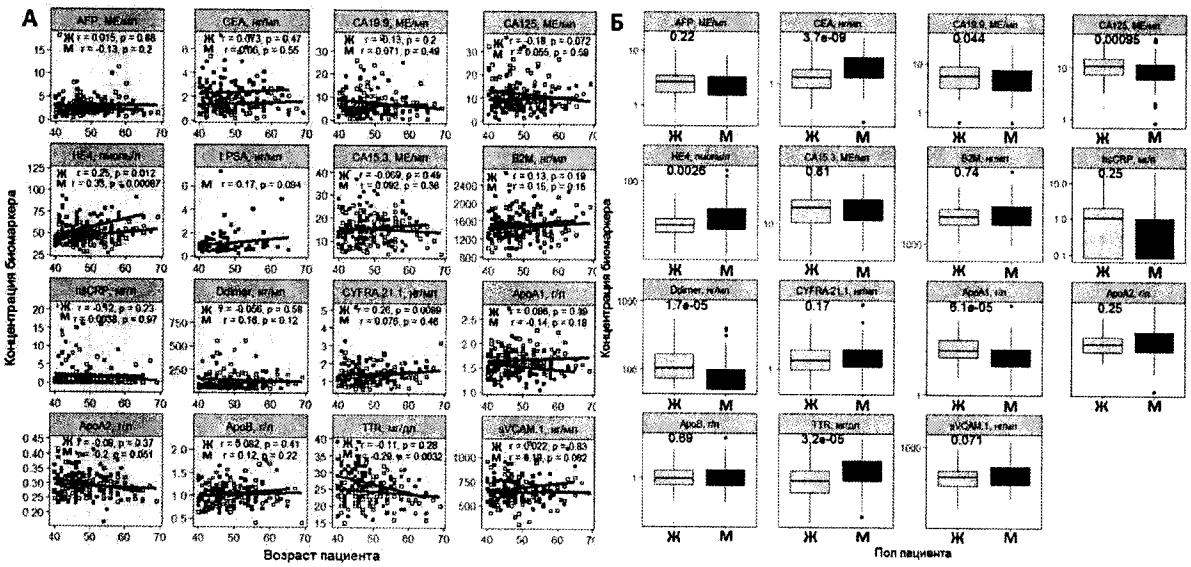
- модуль хранения данных, выполненный с возможностью хранения обучающей и тестовой выборки классификационной модели, связей и весов обученной классификационной модели, записей субъектов с измеренными значениями биомаркеров, включающие записи о пациентах разного пола и возраста;

- модуль обученной классификационной модели, выполненный с возможностью построения и обучения, по меньшей мере, одной классификационной модели для определения наличия заданной патологии по упомянутым маркерам, взятым из модуля хранения данных;

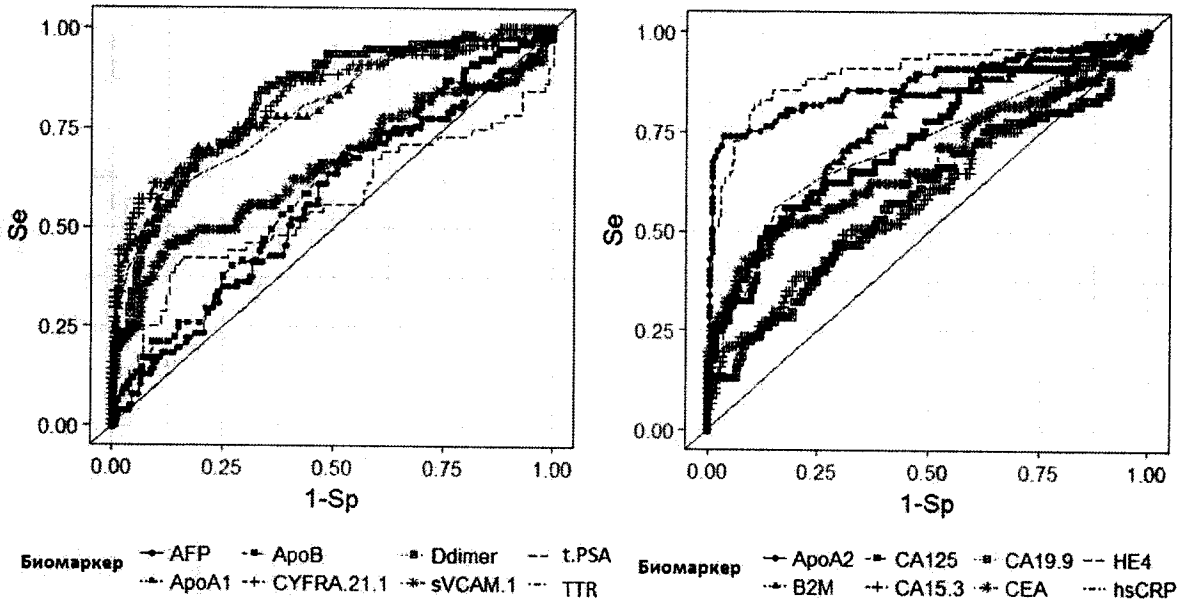
- модуль диагностики, выполненный с возможностью обработки введенных значений биомаркеров субъекта с использованием, по меньшей мере, одной обученной классификационной модели;

- модуль вывода данных, выполненный с возможностью получения данных о высокой или низкой вероятности наличия рака легкого.

СПОСОБ И СИСТЕМА ДЛЯ СКРИНИНГОВОГО ОПРЕДЕЛЕНИЯ ВЕРОЯТНОСТИ НАЛИЧИЯ РАКА ЛЕГКОГО

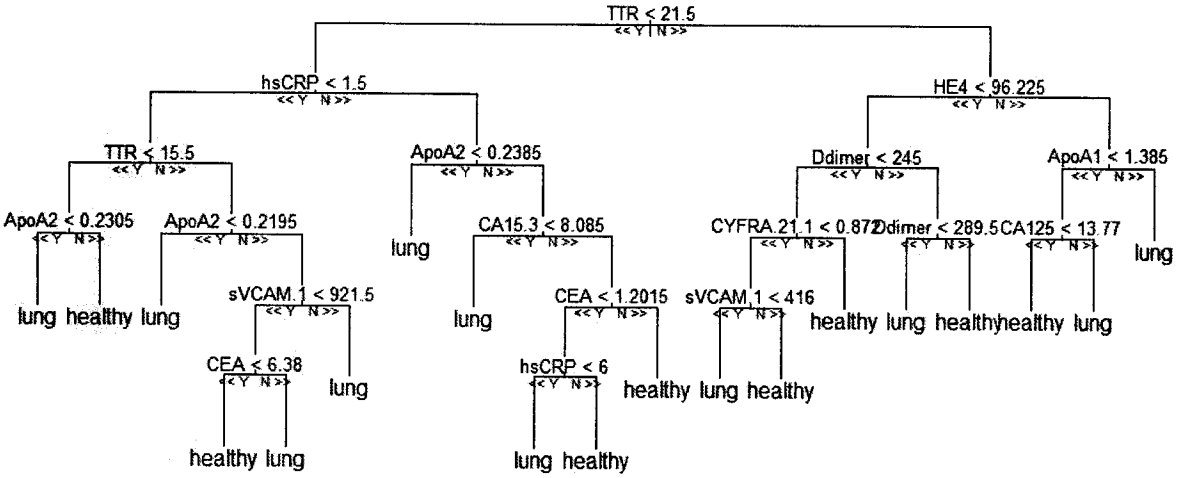


ФИГ.1

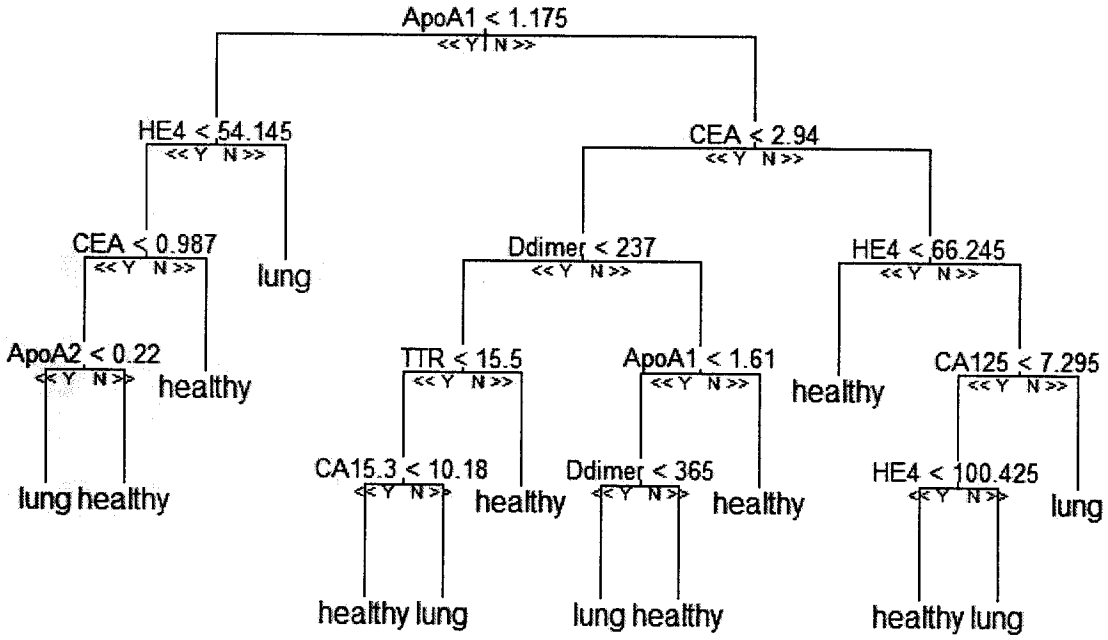


ФИГ.2

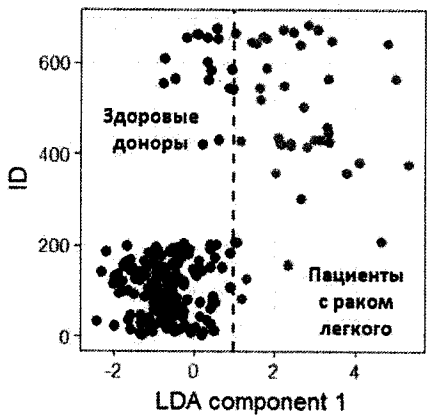
СПОСОБ И СИСТЕМА ДЛЯ СКРИНИНГОВОГО ОПРЕДЕЛЕНИЯ ВЕРОЯТНОСТИ НАЛИЧИЯ РАКА ЛЕГКОГО



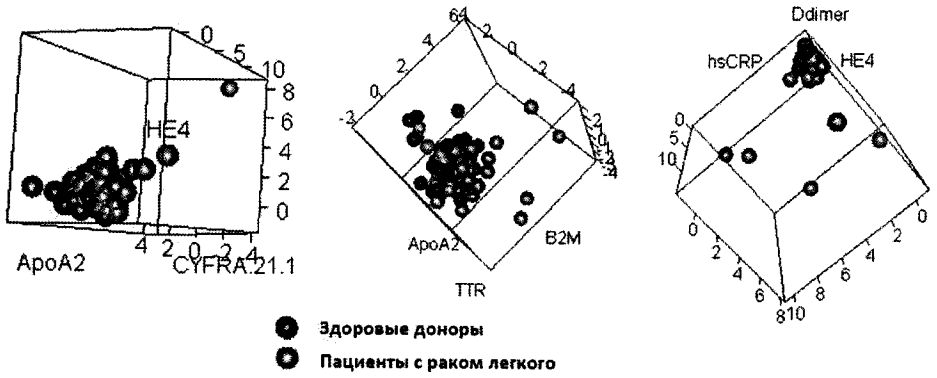
ФИГ.3 А



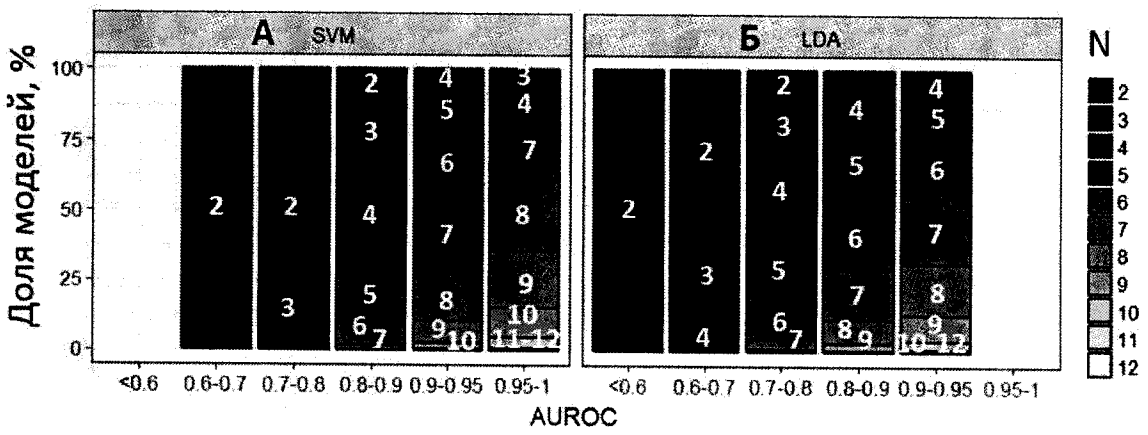
ФИГ.3 Б



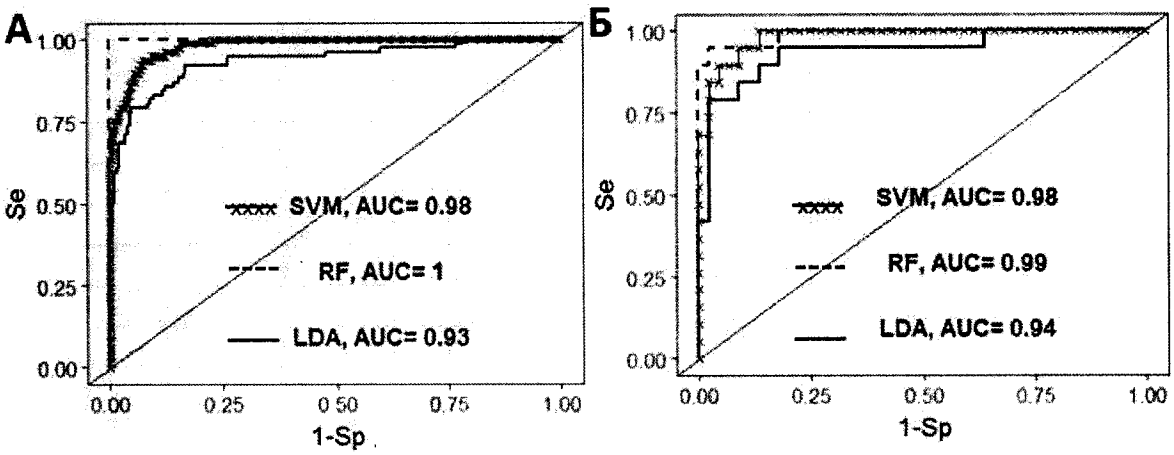
ФИГ.4



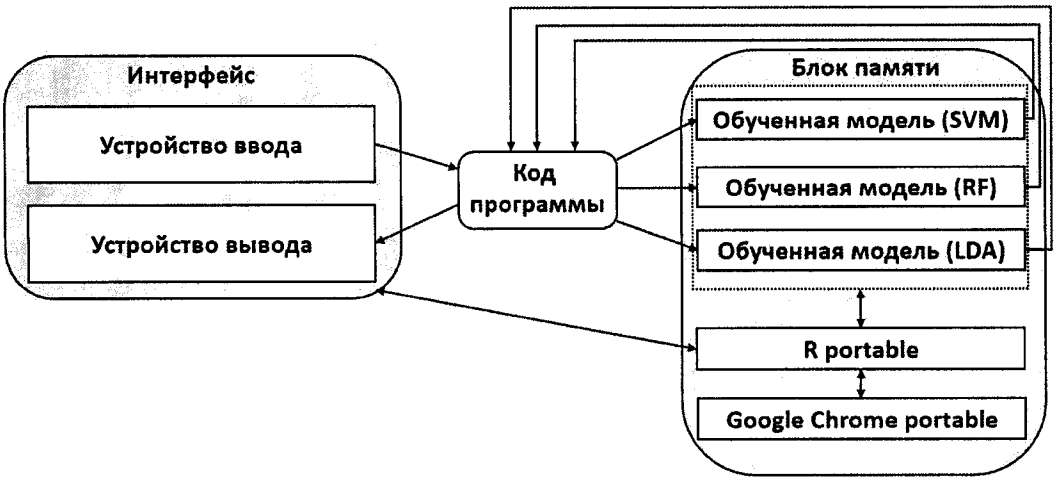
ФИГ.5



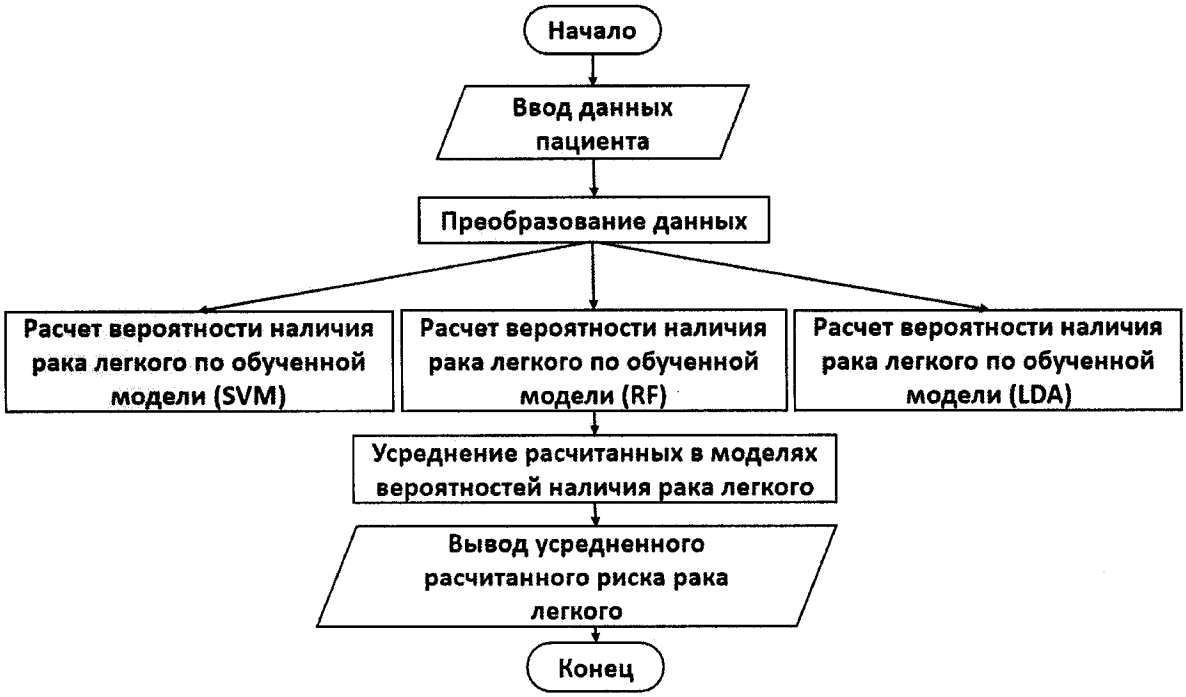
ФИГ.6


А
Se
1.00
0.75
0.50
0.25
0.00
SVM, AUC= 0.98
RF, AUC= 1
LDA, AUC= 0.93
0.00
0.25
0.50
0.75
1.00
1-Sp
Б
Se
1.00
0.75
0.50
0.25
0.00
SVM, AUC= 0.98
RF, AUC= 0.99
LDA, AUC= 0.94
0.00
0.25
0.50
0.75
1.00
1-Sp

ФИГ.7


Интерфейс
Устройство ввода
Устройство вывода
Код программы
Блок памяти
Обученная модель (SVM)
Обученная модель (RF)
Обученная модель (LDA)
R portable
Google Chrome portable

ФИГ.8


Начало
Ввод данных пациента
Преобразование данных
Расчет вероятности наличия рака легкого по обученной модели (SVM)
Расчет вероятности наличия рака легкого по обученной модели (RF)
Расчет вероятности наличия рака легкого по обученной модели (LDA)
Усреднение расчитанных в моделях вероятностей наличия рака легкого
Вывод усредненного расчитанного риска рака легкого
Конец

ФИГ.9

ОТЧЕТ О ПАТЕНТНОМ ПОИСКЕ

(статья 15(3) ЕАПК и правило 42 Патентной инструкции к ЕАПК)

Номер евразийской заявки:

201900375

А. КЛАССИФИКАЦИЯ ПРЕДМЕТА ИЗОБРЕТЕНИЯ:

G16H 50/00 *(2006.01)*
G01N 33/574 *(2006.01)*

Согласно Международной патентной классификации (МПК)

Б. ОБЛАСТЬ ПОИСКА:

Просмотренная документация (система классификации и индексы МПК)
G16H 50/00, A61B 5/00, G01N 33/574

Электронная база данных, использовавшаяся при поиске (название базы и, если, возможно, используемые поисковые термины)

В. ДОКУМЕНТЫ, СЧИТАЮЩИЕСЯ РЕЛЕВАНТНЫМИ

Категория*	Ссылки на документы с указанием, где это возможно, релевантных частей	Относится к пункту №
A	RU 2351936 C1 (ИНСТИТУТ МОЛЕКУЛЯРНОЙ ГЕНЕТИКИ РОССИЙСКОЙ АКАДЕМИИ НАУК) 10.04.2009	1-5
A	RU 2397704 C2 (ГОСУДАРСТВЕННОЕ УЧРЕЖДЕНИЕ НАУЧНО- ИССЛЕДОВАТЕЛЬСКИЙ ИНСТИТУТ ОНКОЛОГИИ ТОМСКОГО НАУЧНОГО ЦЕНТРА СИБИРСКОГО ОТДЕЛЕНИЯ РОССИЙСКОЙ АКАДЕМИИ МЕДИЦИНСКИХ НАУК) 27.08.2010	1-5
A	WO 2013/048292 A2 (КУТУШОВ МИХАИЛ ВЛАДИМИРОВИЧ) 04.04.2013	1-5
A	C. BRAMBILLA et al. Early detection of lung cancer: role of biomarkers, European Respiratory Journal, 2003, 21: Suppl. 39, p. 36s-44s, <DOI: 10.1183/09031936.02.00062002>	1-5

☐ последующие документы указаны в продолжении

* Особые категории ссылочных документов:
«A» - документ, определяющий общий уровень техники
«D» - документ, приведенный в евразийской заявке
«E» - более ранний документ, но опубликованный на дату подачи евразийской заявки или после нее
«O» - документ, относящийся к устному раскрытию, экспонированию и т.д.
"P" - документ, опубликованный до даты подачи евразийской заявки, но после даты испрашиваемого приоритета"

«T» - более поздний документ, опубликованный после даты приоритета и приведенный для понимания изобретения
«X» - документ, имеющий наиболее близкое отношение к предмету поиска, порочащий новизну или изобретательский уровень, взятый в отдельности
«Y» - документ, имеющий наиболее близкое отношение к предмету поиска, порочащий изобретательский уровень в сочетании с другими документами той же категории
«&» - документ, являющийся патентом-аналогом
«L» - документ, приведенный в других целях

Дата проведения патентного поиска: **03/04/2020**

Уполномоченное лицо:
Начальник Управления экспертизы _____ Д.Ю. Рогожин