(19) Евразийское патентное ведомство

(21) 202090713 (13) A2

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОЙ ЗАЯВКЕ

(43) Дата публикации заявки
2020.10.30

(22) Дата подачи заявки
2020.04.10

(51) Int. Cl. ***G16H 50/20*** (2018.01)
G16H 50/30 (2018.01)

(54) СПОСОБ СКРИНИНГОВОГО ОПРЕДЕЛЕНИЯ ВЕРОЯТНОСТИ НАЛИЧИЯ РАКА МОЛОЧНОЙ ЖЕЛЕЗЫ

(31) **2019111094**
(32) **2019.04.12**
(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, Васильев Иван Алексеевич, 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 молочной железы или выявления данного онкологического заболевания на ранней стадии. Скрининговое определение вероятности наличия рака молочной железы основано на измерении уровня биомаркеров в образце биологической жидкости, полученном у субъекта: CYFRA.21.1, ApoA2, Ddimer, HE4, B2M, ApoA1, sVCAM.1, CA125, CA15.3, TTR, hsCRP, CEA, с последующей обработкой совокупности полученных данных с использованием по меньшей мере одной классификационной модели, обученной для определения вероятности наличия рака молочной железы.

202090713 A2

202090713 A2

МПК: G16H50/20, G16H50/30

СПОСОБ СКРИНИНГОВОГО ОПРЕДЕЛЕНИЯ ВЕРОЯТНОСТИ НАЛИЧИЯ РАКА МОЛОЧНОЙ ЖЕЛЕЗЫ

Область техники, к которой относится изобретение

Изобретение относится к области медицины, а именно онкологии, и может быть использовано для скринингового определения вероятности наличия рака молочной железы (РМЖ) 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 по России в 2016г. составил 408,6 чел. на 100000 населения. Средний показатель смертности – 201,6 чел. на 100000 населения. Онкологическая заболеваемость растет во всем мире. За последние 10 лет она увеличилась более чем на 20%.

В случае развития злокачественного заболевания, стадия, на которой онкологический процесс будет выявлен, является одним из определяющих факторов, обуславливающих продолжительность жизни пациента.

Рак молочной железы (РМЖ) является лидирующим онкологическим заболеванием во всех странах мира. Ежегодно регистрируется более миллиона случаев данного заболевания. При этом в случае обнаружения заболевания на ранней стадии в подавляющем большинстве случаев возможно полное излечение.

Для обнаружения рака молочной железы известно применение следующих инструментальных методов диагностики – маммографии, дуктографии, ультразвукового исследования (УЗИ), магнитно-резонансной томографии, позитронно-эмиссионной томографии и др.

Метод УЗИ, наиболее распространенный у женщин возраста до 35 лет, позволяет отличить более плотную ткань опухоли от окружающей нормальной ткани. Однако в процессе диагностики участки жировой инволюции могут быть ошибочно приняты за патологические структуры. Кроме того, УЗИ не визуализирует микрокальцинаты, часто встречающиеся при злокачественных новообразованиях.

Основным на сегодняшний день методом диагностики злокачественных новообразований молочной железы является маммография, которая позволяет с

достоверностью до 95% выявить новообразования молочной железы размером более 10 мм в диаметре.

Однако несмотря на эффективность рентгенологического метода, в ряде случаев разрешающая способность маммографии резко снижается, например, при выраженных воспалительных изменениях, диффузных формах мастопатии, отеке железы и фоновых заболеваниях типа фиброаденоматоза.

Общим ограничением инструментальных методов диагностики является неоднозначность интерпретации результатов, связанная с многообразием индивидуальных особенностей строения и морфологией молочной железы.

В качестве альтернативы инструментальным диагностическим методам 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 могут быть использованы различные антигены, протеины и метаболиты, секретируемые злокачественными клетками или образующиеся в процессе их гибели. На текущий момент не существует рекомендаций по использованию биомаркеров для диагностики РМЖ. Одними из наиболее перспективных кандидатов являются CEA (раковый эмбриональный антиген), CYFRA 21-1 (фрагмент цитокератина 19) и CA15-3 (раковый антиген 15-3), однако их использование для диагностики РМЖ ограниченно ввиду недостаточной чувствительности и специфичности (Kazarian et al., Testing breast cancer serum biomarkers for early detection and prognosis in pre-diagnosis samples. Br J Cancer. 2017 Feb 14; 116(4): 501–508). Использование мультиплексных диагностических методов, подразумевающих оценку риска наличия заболевания на основе измерений нескольких биомаркеров, позволяет преодолеть данную проблему и достичь более достоверных результатов.

Из международной заявки WO2013062931 известно определение метастазов рака молочной железы по обнаружению матричных РНК специфических биомаркеров, циркулирующих в периферической сыворотке крови, костного мозга или лимфатических узлов – CEA, CA15-3, PIP, hMAM и HER2.

В патенте US6670141 представлена панель биомаркеров для диагностики и лечения рака молочной железы. При этом проводились исследования слюны здоровых женщин, женщин с доброкачественными поражениями молочной железы и женщин с диагнозом рака молочной железы. Было выявлено, что уровни маркеров c-ErbB-2 (ERB) и CA 15-3 у больных раком значительно выше, чем у здоровых женщин и женщин с

доброкачественными опухолями, а уровень белка p53, напротив, выше в контрольной группе здоровых женщин.

Наборы биомаркеров для диагностики РМЖ предложены также в заявках US20160282351, US20150024960, WO2010017515.

В заявке WO2005113835 были определены потенциальные биомаркеры рака молочной железы, идентифицированные в образцах протокового лаважа от отдельных женщин с высоким риском развития РМЖ, среди которых выявлены в т.ч. Apoliprotein A-I (ApoA-I), Apoliprotein A-II (ApoA-II). Было обнаружено, что более низкий, чем обычно, уровень экспрессии данных маркеров или комбинации маркеров коррелирует с раком молочной железы у пациента.

Патентная заявка WO2005083440 описывает способ диагностирования рака яичников, рака молочной железы и рака прямой кишки на основе одновременной идентификации множества биомаркеров, продукция которых в организме резко возрастает при онкологических заболеваниях. Среди данных биомаркеров предложены в т.ч. лептин, пролактин, ферменты химотрипсиного ряда, калликреины, онкомаркеры CA125, CA15-3, CA19-9, MUC1, OVX1, РЭА, M-CSF, OPN и IGF-II, простатин, CA54-61, CA72, HMFG2, интерлейкины IL-6, IL-10, LSA, M-CSF, NB70K, PLAP, TAG72, факторы TNF, TPA, UGTF, VEGF, CLDN3, NOTCH3, E2F и др. При этом для диагностики РМЖ выделены биомаркеры лептина, пролактина, OPN и IGF-I.

Из патента RU2599890 известен способ обнаружения и мониторинга терапии рака молочной железы и рака яичников на основе определения концентраций опухолеассоциированных антигенов: AFP, hCG, CEA, CA125, CA15-3 и CA19-9 в образце сыворотки человека.

Из публикации (Kim et al., The multiplex bead array approach to identifying serum biomarkers associated with breast cancer. Breast Cancer Research 2009, 11:R22, doi:10.1186/bcr2247) (прототип) известен способ оценки риска возникновения различных видов рака, в т.ч. РМЖ, по измеренным в сыворотке белковым биомаркерам - AFP, CEA, CA19-9, CA125, PSA, ApoA1, ApoA2, TTR, B2M, IL-6, CRP, PAI-1.

Несмотря на высокую дискриминационную способность предложенной в прототипе модели, необходима ее валидация и адаптация для различных популяций обследуемых, что связано с межпопуляционными различиями в молекулярных механизмах канцерогенеза. Так, были выявлены различия в экспрессии генов, характере соматических мутаций и паттернах метилирования ДНК из образцов опухолей, взятых у пациенток европеоидной и черной расы (Huo et al., Comparison of Breast Cancer Molecular Features and Survival by

African and European Ancestry in The Cancer Genome Atlas. JAMA Oncol. 2017 Dec 1;3(12):1654-1662). В работе отмечено, что подобные различия могут влиять на встречаемость отдельных подтипов рака молочной железы и прогноз заболевания.

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

Раскрытие изобретения

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

Достигаемым техническим результатом является повышение точности скринингового выявления наличия рака у пациенток европеоидной популяции, причем уже на ранних стадиях его развития, посредством биостатистической обработки результатов анализа фракции сыворотки и плазмы крови с определением концентрации комплексной группы биомаркеров.

Технический результат достигается посредством реализации способа скринингового определения вероятности наличия РМЖ, включающего измерение уровня биомаркеров в образце биологической жидкости (например, плазмы или цельной крови, мочи, мокроты), полученном у субъекта: HE4, ApoA2, CYFRA.21.1, Ddimer, ApoA1, TTR, B2M, CA125, hsCRP, CEA, sVCAM.1, CA15.3, с последующей обработкой совокупности полученных значений биомаркеров с использованием, по меньшей мере, одной классификационной модели, обученной для определения высокой или низкой вероятности наличия РМЖ.

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 пациентках разного 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 РМЖ по итогам обработки измеренных данных биомаркеров субъекта.

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

Технический результат достигается посредством реализации системы скринингового определения вероятности наличия РМЖ, включающей

- 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 HE4, ApoA2, CYFRA.21.1, Ddimer, ApoA1, TTR, B2M, CA125, hsCRP, CEA, sVCAM.1, CA15.3, включающие записи о пациентках разного 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 РМЖ.

Точность заявляемого мультиплексного метода диагностики РМЖ обеспечивается за счет использования комплекса из 12 биомаркеров, а также за счет использования нескольких классификационных моделей с последующим усреднением модельных результатов.

Краткое описание чертежей

Изобретение поясняется чертежами, где:

На фиг.1 представлена диаграмма рассеяния «возраст пациентки - 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.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 пациентки с РМЖ) при помощи линейного дискриминантного анализа по 12 биомаркерам;

На фиг.5 - Примеры 3-мерных проекций разделения объединенной популяции пациенток на 2 класса (здоровые доноры и пациентки с РМЖ) 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 РМЖ на основе данных пациентки.

На фиг.9 - Алгоритм оценки вероятности наличия РМЖ на основе данных пациентки.

Осуществление изобретения

Исходная группа биомаркеров, используемая в диагностическом тесте на определение вероятности наличия РМЖ 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 пациенток с

РМЖ. Обученная модель может быть использована для оценки риска наличия заболевания у пациентки на основе показателей ее 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 (104 женщины, средний возраст 50 лет) и пациенток с РМЖ (86 женщин, средний возраст 63 года).

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 возраста на показатели биомаркеров (Фиг.1). По итогам проведения исследования на данном этапе был сделан вывод об отсутствии значимой корреляции между возрастом пациенток и показателями большинства биомаркеров (за исключением CYFRA.21.1 и HE4).

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

Таблица 1. Сравнение показателей биомаркеров в плазме крови здоровых доноров и пациенток с РМЖ.

Биомаркер	Здоровые доноры	Пациентки с РМЖ	P-значения[1]
AFP, МЕ/мл	3.07	3.19	0.2
ApoA1, г/л	1.66	1.45	2e-08 ***
ApoA2, г/л	0.29	0.24	1e-12 ***
ApoB, г/л	1.01	1.01	0.5
B2M, нг/мл	1464.48	2009.78	4e-09 ***
CA125, МЕ/мл	11.83	29.73	4e-05 ***
CA15.3, МЕ/мл	14.88	32.59	6e-05 ***
CA19.9, МЕ/мл	7.30	11.24	0.1
CEA, нг/мл	1.40	2.97	0.01 *
CYFRA.21.1, нг/мл	1.32	3.93	7e-13 ***
Ddimer, нг/мл	140.13	402.03	3e-10 ***

HE4, пмоль/л	48.01	76.64	3e-09 ***
hsCRP, мг/л	1.85	5.23	6e-04 ***
sVCAM.1, нг/мл	639.25	795.70	4e-07 ***
TTR, мг/дл	24.29	21.14	1e-05 ***

[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 и ApoB между здоровыми добровольцами и пациентками с РМЖ. Также в рамках данного исследования отмечено значимое различие в концентрациях CA15.3 и Ddimer, не включенных в прототип.

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
CYFRA.21.1	44	84	70	0.84
ApoA2	43	72	74	0.79
Ddimer	39	70	76	0.77
HE4	41	77	67	0.77
B2M	42	75	70	0.75
ApoA1	51	78	57	0.72
sVCAM.1	46	78	63	0.71
CA125	36	42	85	0.67
CA15.3	57	95	35	0.67
TTR	56	88	38	0.65
hsCRP	44	84	43	0.64
CEA	45	67	50	0.60
CA19.9	42	52	67	0.59
AFP	45	43	78	0.57
ApoB	45	88	19	0.49

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 (**CYFRA.21.1, ApoA2, Ddimer, HE4, B2M, ApoA1, sVCAM.1, CA125, CA15.3, TTR, hsCRP, CEA**).

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 пациентках с РМЖ, 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 пациентка причисляется к одной из групп (здоровые доноры или пациентки с РМЖ), 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 пациенток с РМЖ. 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
CEA	-6.25E-02
CA125	-9.33E-04
HE4	-2.25E-03
CA15.3	3.41E-03
B2M	9.14E-05
hsCRP	2.47E-02
Ddimer	5.04E-04
CYFRA.21.1	1.08E-01
ApoA1	-4.73E-01
ApoA2	-1.68E+01
TTR	3.75E-02
sVCAM.1	2.00E-03

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 $F(x) = \text{sign}(\langle w, x \rangle + b)$, где w — нормальный вектор к разделяющей гиперплоскости, b — вспомогательный параметр, а функция 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 пациенток, 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 пациентках с РМЖ, 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, наибольшей 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 пациентках с РМЖ, 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	70	91	81	50	0.88
SVM	80	93	87	50	0.93
80% данных было использовано для обучения модели, 20% - для валидации					
RF	75	100	88	50	0.95
LDA	70	90	81	50	0.81
SVM	85	100	93	50	0.92

Финальные классификационные модели представляют собой обученные алгоритмы, позволяющие предсказать вероятность наличия РМЖ на основании экспериментальных измерений биомаркеров пациенток.

Финальное решение - определение вероятности наличия РМЖ, рассчитывается как медиана значений вероятностей наличия РМЖ, рассчитанных в 3 классификационных моделях (RF, LDA SVM), обученных на всей выборке пациенток (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 конкретной пациентки (результаты измерения биомаркеров) рассчитывать вероятность наличия у нее РМЖ. Блок-схема реализации изобретения представлена на фиг. 8.

Компьютерно-реализуемая система состоит из (1) интерфейса, включающего устройство ввода данных пациентки (результаты измерений биомаркеров) и вывода результатов расчета (вероятность наличия РМЖ); (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 РМЖ на основе данных пациентки представлен на фиг.9.

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

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

Ниже представлены клинические примеры применения способа.

Пример 1. Больная И., 63 года.

Пациентке было предложено принять участие в программе Онкопоиска.

Пациентка (724) обследована в рамках программы. Получены следующие результаты:

AFP 2,7 МЕ/мл, CEA 50,4 нг/мл, CA 19-9 77,8 МЕ/мл, CA15.3 824,4 МЕ/мл, CA125 820,9 МЕ/мл, B2M 3261 нг/мл, hsCRP 2 нг/мл, HE4 138,8 пмоль/л, Ddimer 473,0 нг/мл, CYFRA.21.1 19,92 нг/мл, Apo A1 1,01 г/л, Apo A2 0,16 г/л, Apo B 0,9 г/л, TTR (prealb) 15,0 мг/дл, sVCAM.1 998 нг/мл, Rantes 33263 пг/мл, VEGFR1 151 пг/мл, LRG-1 136201 нг/мл, Apo A4 37,8 мкг/мл.

При обработке полученных результатов заявляемым способом выявлена вероятность РМЖ, значительно превышающая пороговое значение, составляющее 50% (Таблица 4): модель RF – 99.8%, модель LDA – 99.1%, модель SVM – 94.5%, усредненное итоговое значение вероятности наличия РМЖ по трем моделям составило 97.8%.

Пациентка приглашена на обследование.

При осмотре:

Периферические лимфатические узлы не увеличены. Молочные железы симметричны. Соски и ареолы не изменены. В обеих молочных железах узловые образования не определяются, ткань молочных желез повышенной плотности. Кожных симптомов нет.

Пациентка направлена на маммографию с томосинтезом.

На маммографии молочные железы повышенной маммографической плотности (4). В ткани обеих молочных желез остаточные явления фиброзно-кистозной мастопатии с выраженным фиброзным компонентом. В верхне-наружном квадранте правой молочной железы определяется участок со скоплением микрокальцинатов, площадью до 2 см.

Под контролем УЗИ выполнена пункция подозрительного в отношении рака участка правой молочной железы.

Цитологическое заключение: клетки рака.

Пациентке проведено хирургическое лечение в объеме радикальной резекции правой молочной железы. Гистологическое заключение: В ткани сектора молочной железы опухолевый узел размером 0,8 см в диаметре. Имеет строение инфильтративного протокового рака II степени злокачественности, с наличием структур протокового рака in situ.

В 18 исследованных лимфоузлах- метастазы рака не выявлены.

Иммуногистохимическое исследование: ER (Рецепторы эстрогенов) 105 Н-баллов, PR (рецепторы прогестерона) - 50 Н-баллов. Her2neu 1+.

При обследовании данных за отдаленные метастазы не выявлено.

Диагноз: Рак правой молочной железы T1N0M0.

Пример 2. Больная К., 72 года.

Пациентке было предложено принять участие в программе Онкопоиска.

Пациентка (659) обследована в рамках программы. Получены следующие результаты:

AFP 4,14 МЕ/мл, CEA 1,68 нг/мл, CA 19-9 14.92 МЕ/мл, CA15.3 57,29 МЕ/мл, CA125 15,11 МЕ/мл, B2M 3527 нг/мл, hsCRP 32 нг/мл, HE4 143,3 пмоль/л, Ddimer 466,0 нг/мл, CYFRA.21.1 4,79 нг/мл, Apo A1 1,65 г/л, Apo A2 0,344 г/л, Apo B 2,04 г/л, TTR (prealb) 26,0 мг/дл, sVCAM.1 1278 нг/мл, Rantes 59201 пг/мл, VEGFR1 91 пг/мл, LRG-1 98732 нг/мл, Apo A4 32,5 мкг/мл.

При обработке полученных результатов заявляемым способом выявлена высокая вероятность рака молочной железы.

При обработке полученных результатов заявляемым способом выявлена вероятность РМЖ значительно превышающая пороговое значение, составляющее 50% (Таблица 4): модель RF – 96.6%, модель LDA – 81.1%, модель SVM – 94.8%, усредненное итоговое значение вероятности наличия РМЖ по трем моделям составило 90.8%.

Пациентка приглашена на обследование.

При осмотре периферические л/узлы не увеличены. В ткани обеих молочных желез опухолевые образования не определяются.

Ранее в прошлом году пациентка была обследована, выполнена маммография. По данным заключения: инволютивные изменения в ткани молочных желез.

Выполнена маммография с томосинтезом. В верхненаружном квадранте правой молочной железы определяется скопление микрокальцинатов на площади 2 см. Выполнена Cor-биопсия этого участка ткани молочной железы под контролем прицельной маммографии.

Гистологическое заключение: картина инвазивного без признаков специфичности рака молочной железы G2, трабекулярно-солидного строения. Иммуногистохимическое исследование: рецепторы эстрогенов 8 баллов, рецепторы прогестерона 8 баллов, Ki 67 4%, Her2neu – 0.

Установлен диагноз рака молочной железы T1N0M0, люминальный А.

Оперативное лечение в объеме радикальной резекции правой молочной железы.

Диагноз подтвержден гистологически. Опухоль размерами 15мм, без признаков сосудистой и периневральной инвазии. В 17 исследованных лимфатических узлах без признаков метастазирования.

ФОРМУЛА ИЗОБРЕТЕНИЯ

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

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:

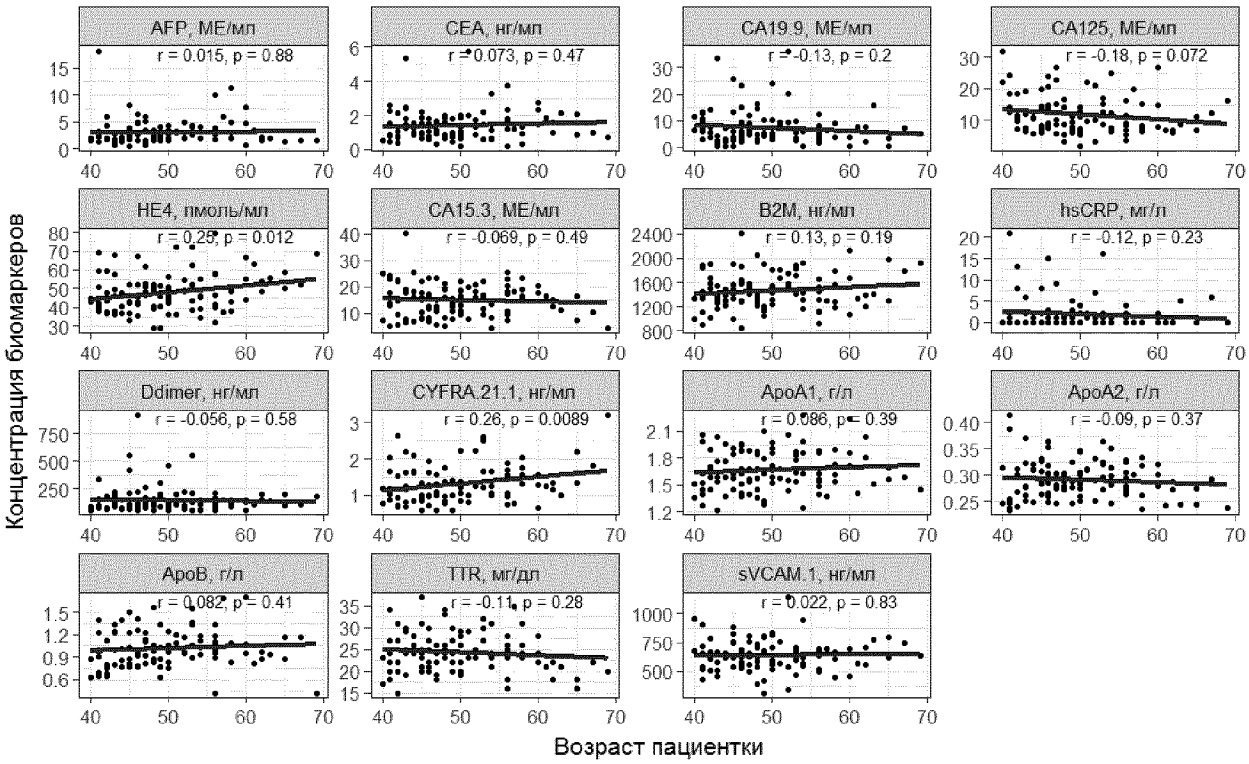
- формируют обучающую и тестовую выборку записей субъектов с измеренными значениями биомаркеров CYFRA.21.1, ApoA2, Ddimer, HE4, B2M, ApoA1, sVCAM.1, CA125, CA15.3, TTR, hsCRP, CEA, включающие записи о пациентках разного возраста;

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

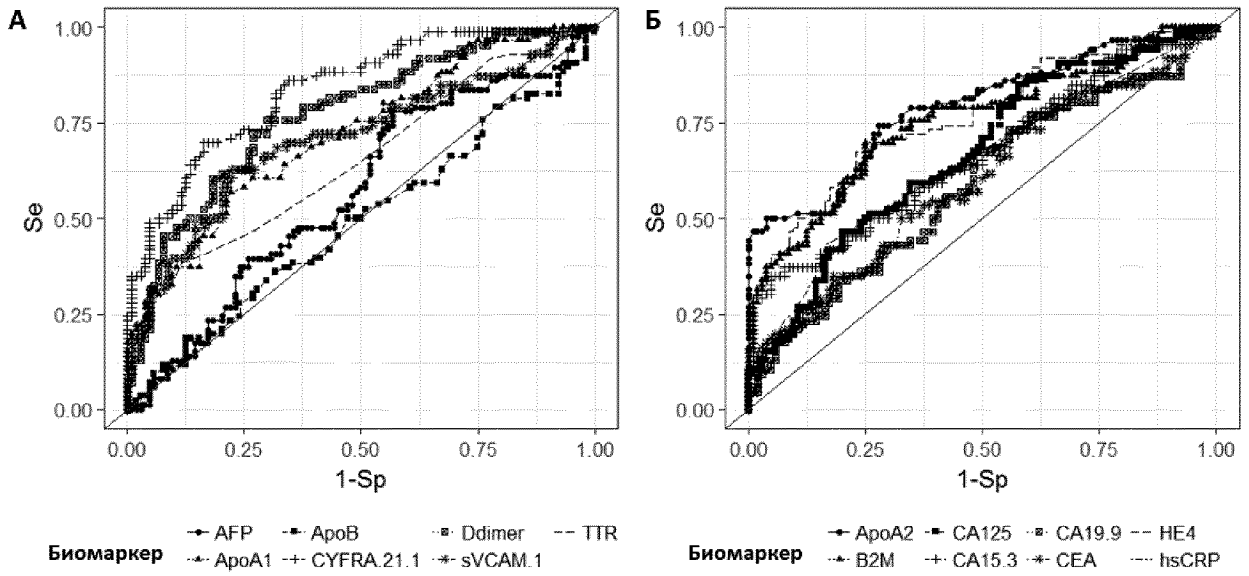
- сохраняют связи и веса обученной классификационной модели, для последующего определения вероятности наличия РМЖ по итогам обработки измеренных данных биомаркеров субъекта.

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

СПОСОБ СКРИНИНГОВОГО ОПРЕДЕЛЕНИЯ ВЕРОЯТНОСТИ НАЛИЧИЯ РАКА МОЛОЧНОЙ ЖЕЛЕЗЫ

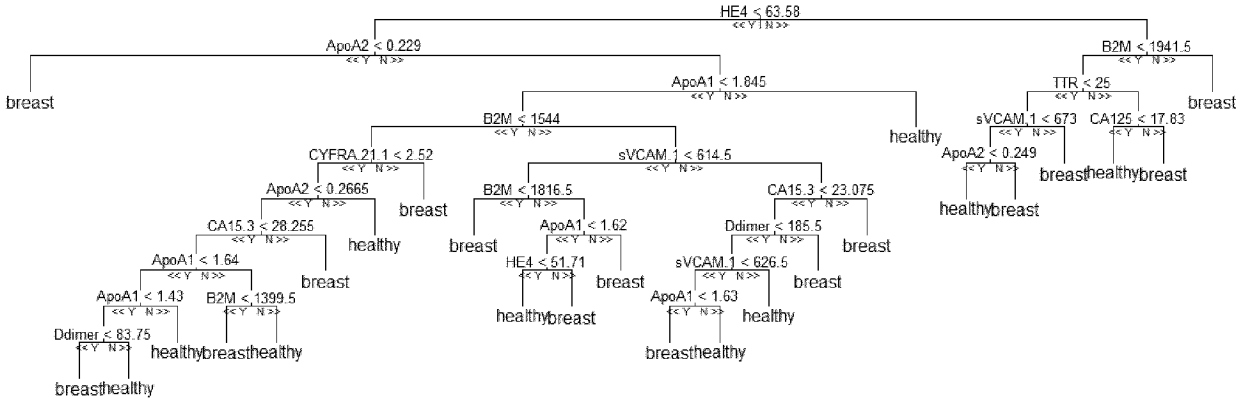


ФИГ.1

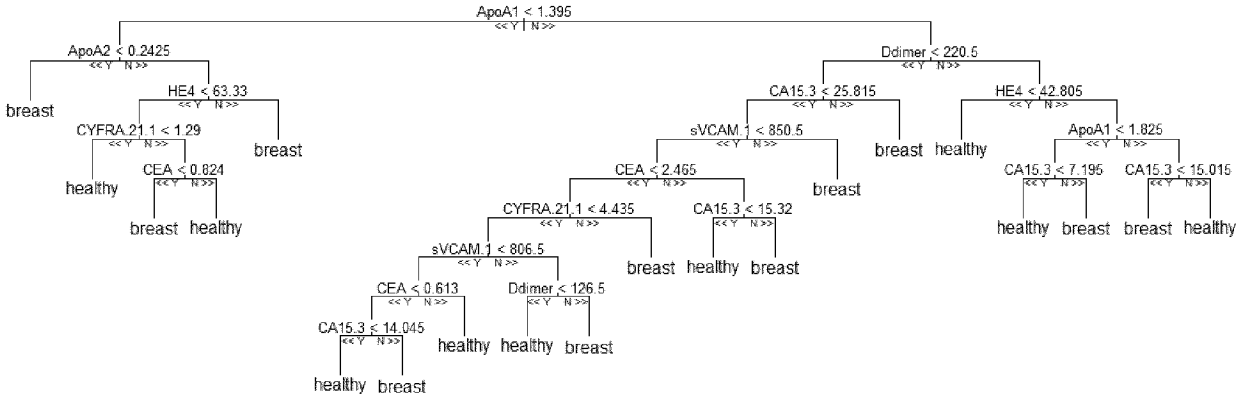


ФИГ.2

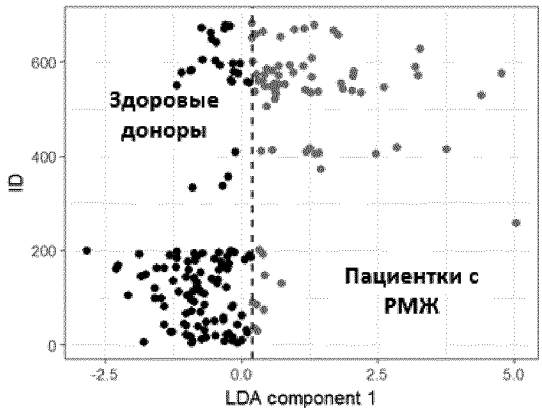
СПОСОБ СКРИНИНГОВОГО ОПРЕДЕЛЕНИЯ ВЕРОЯТНОСТИ НАЛИЧИЯ РАКА МОЛОЧНОЙ ЖЕЛЕЗЫ



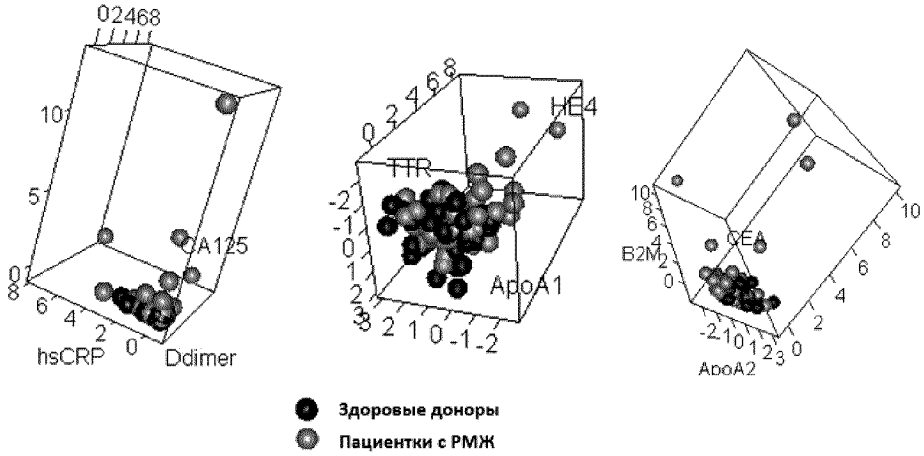
ФИГ.3 А



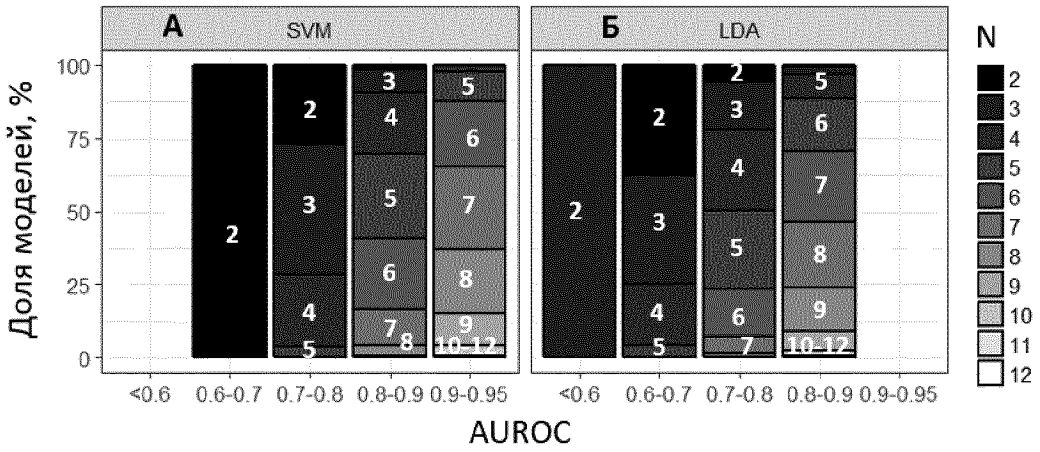
ФИГ.3 Б



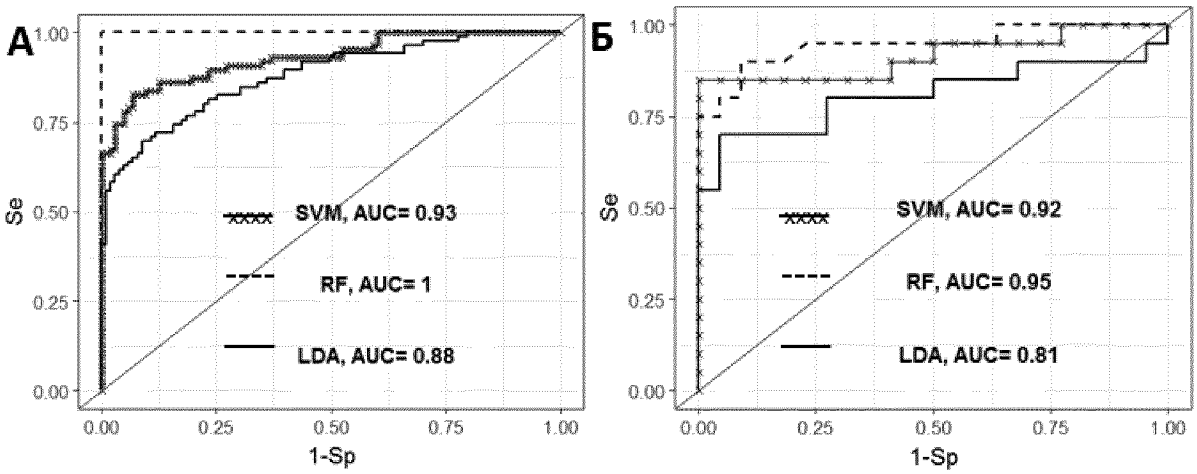
ФИГ.4



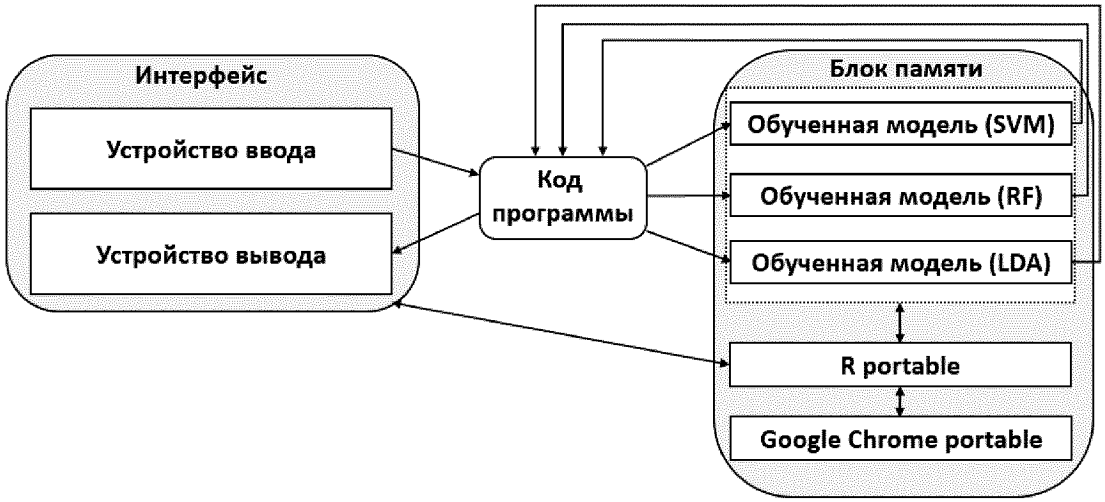
ФИГ.5



ФИГ.6

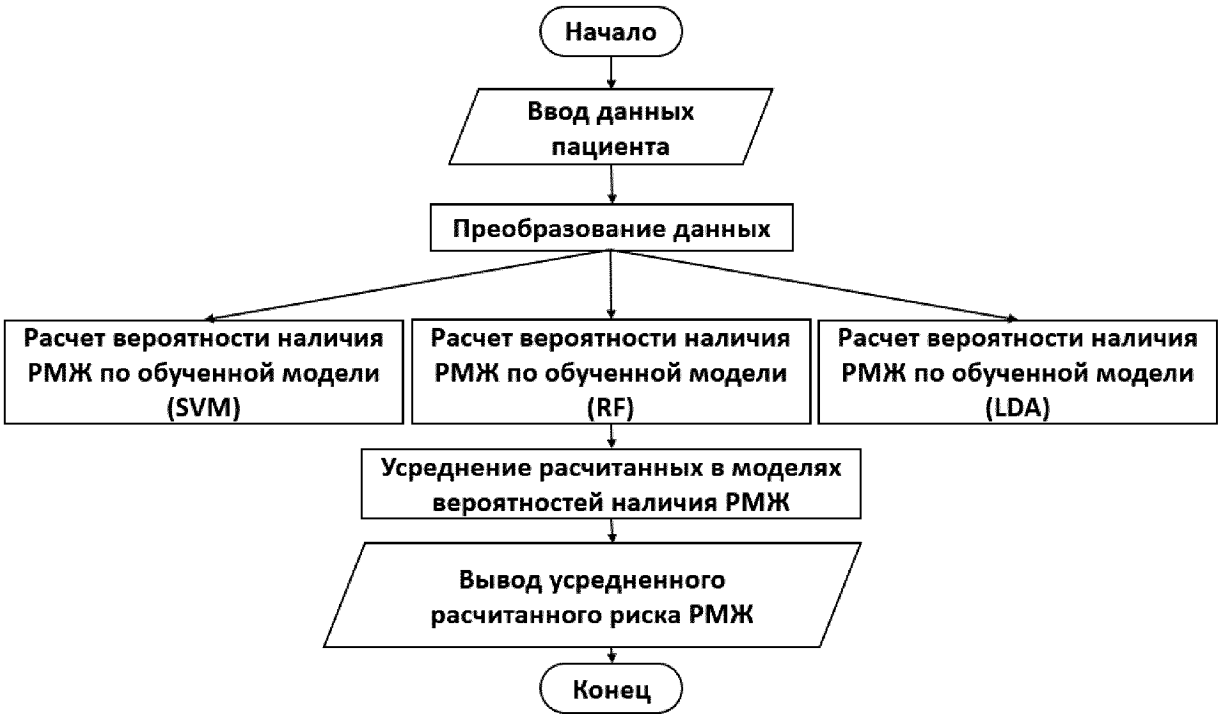

А
Se
1-Sp
SVM, AUC= 0.93
RF, AUC= 1
LDA, AUC= 0.88
Б
Se
1-Sp
SVM, AUC= 0.92
RF, AUC= 0.95
LDA, AUC= 0.81

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

СПОСОБ СКРИНИНГОВОГО ОПРЕДЕЛЕНИЯ ВЕРОЯТНОСТИ НАЛИЧИЯ РАКА МОЛОЧНОЙ ЖЕЛЕЗЫ



ФИГ.9