Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»
при Вычислительном центре им. А. А. Дородницына РАН

Целых Влада Руслановна

Статистические обоснования информационного анализа электрокардиосигналов для диагностики заболеваний внутренних органов

010958 — Прикладная информатика

Магистерская диссертация

Научный руководитель:
ст.н.с. ВЦ РАН, д.ф.-м.н.
Воронцов Константин Вячеславович

Москва
2015 г.

Содержание

Аннотация

Технология информационного анализа электрокардиосигналов, основанная на теории информационной функции сердца, уже более 10 лет применяется во врачебной практике для определения рисков наиболее опасных и распространенных заболеваний внутренних органов. За это время накоплена выборка, содержащая более 20 тысяч обследований как больных различными заболеваниями, так и здоровых людей. На основе этих наблюдений в данной работе приводится статистическое обоснование самой возможности диагностики многих заболеваний внутренних органов по одной электрокардиограмме, сравнивается точность диагностики 18 заболеваний с помощью линейных моделей классификации и случайного леса. Эксперименты подтверждают высокие уровни чувствительности и специфичности для всех заболеваний.

1 Введение

Из физиологии сердца человека известно, что электрический, магнитный и гидродинамический импульсы, генерируемые сердцем во время его работы, являются источником важной информации о состоянии сердца и системы регуляции его функций. Электрофизиологические методы исследования и, в первую очередь, электрокардиография, получившая приоритетное развитие, в настоящее время играют важнейшую роль в современной кардиологии, в научной и практической медицине. Они позволяют достаточно глубоко оценить состояние миокарда и функций сердца.

Опыт изучения вариабельности сердечного ритма (ВСР) на основе длительной регистрации электрокардиограммы свидетельствует о том, что электрокардиоимпульсы могут быть носителями информации также о состоянии системы регуляции основных функций организма в норме, при различных заболеваниях и в условиях воздействия на человека экстремальных факторов профессиональной деятельности и среды обитания [1, 2]. Сердце генерирует импульсы электрической, магнитной и гидродинамической природы под влиянием сложного комплекса компонентов системы регуляции. Эти импульсы распространяются в объёме всего организма и имеют свойства сигналов. Вариации сердечного ритма носят иррегулярный характер, однако изолированное сердце вне организма человека генерирует импульсы без какой-либо вариабельности.

На основе этих наблюдений в [3] предлагается *теория информационной функции*

сердца и обосновывается роль сердца как информационного органа. Предположение о том, что в организме существуют механизмы передачи сигналов, аналогичные амплитудной и частотной модуляции в теории сигналов и связи, приводит к идее исследования вариабельности не только R-R-интервалов, но и R-амплитуд. Поиск способов демодуляции этих сигналов и дешифровки содержащейся в них информации о заболеваниях привёл к созданию *технологии информационного анализа электрокардиосигналов* [3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. Исследования в этом направлении проводились на базе Военно-медицинской академии (г. Санкт-Петербург), Российской медицинской академии последипломного образования, Российской академии космонавтики имени К. Э. Циолковского (ЦНИБИП), Государственного института усовершенствования врачей МО РФ и 2-го Центрального военного клинического госпиталя им. П. В. Мандрыка МО РФ, ныне Медицинского учебно-научного клинического центра им. П. В. Мандрыка.

В настоящее время технология информационного анализа электрокардиосигналов реализована в диагностической системе «Скринфакс» [3]. Она позволяет диагностировать по одной электрокардиограмме более 30 различных заболеваний внутренних органов, не ограничиваясь заболеваниями сердечно-сосудистой системы. За 10 лет врачебного применения накоплена обучающая выборка более 20 тысяч прецедентов — записей электрокардиограмм и соответствующих им диагнозов.

Технология информационного анализа электрокардиосигналов основана на преобразовании каждой электрокардиограммы сначала в последовательность интервалов и амплитуд кардиоциклов, затем в символьную последовательность — кодограмму и, наконец, в числовой вектор — признаковое описание фиксированной размерности, что позволяет строить диагностические правила по обучающей выборке методами машинного обучения.

В данной работе приводятся статистические обоснования отдельных этапов технологии информационного анализа электрокардиосигналов. По представительным выборкам здоровых людей и больных 18 различными заболеваниями внутренних органов проверяются статистические гипотезы о неслучайном характере вариаций интервалов и амплитуд кардиоциклов и о взаимосвязи этих вариаций с заболеваниями.

Целью работы является статистическое обоснование информационного анализа электрокардиосигналов, повышение точности диагностики путём оптимизации моделей классификации и исследование их обобщающей способности.

2 Обзор литературы

2.1 Методы ВСР-анализа

Анализ вариабельности сердечного ритма (ВСР) основан на определении последовательности R-R-интервалов электрокардиограммы. Методы оценки ВСР можно разбить на 3 группы [2, 13, 14]:

- *методы временной области* — включают в себя показатели, получаемые непосредственно из ряда интервалов или из ряда их разностей (среднее значение, мода, стандартное отклонение и т.д.);
- *методы частотной области* — основаны на спектральном разложении ряда интервалов (мощность сверхнизкочастотных (VLF), низкочастотных (LF) и высокочастотных колебаний (HF), отношение LF/HF и т.д.);
- *методы нелинейного анализа* — имеют сложную интерпретацию, отражают нелинейную динамику ритма сердца (корреляционная размерность, длины полуосей эллипса скаттерограммы, приближенная энтропия, показатели символьной динамики и т.д.).

Первые две группы методов включают в себя стандартные показатели, широко использующиеся на практике [2]. Последняя группа методов продолжает пополняться, ее основные показатели описаны в [14].

Диагностическая ценность методов символьной динамики была впервые установлена в 1995 году [15]. Статистические показатели рассчитывались из анализа символьных последовательностей, полученных с помощью кодирования ряда R-R-интервалов. Было предложено два способа 4-буквенного кодирования: по величине отклонения текущего значения от среднего значения интервалов, а также на основе разности между соседними величинами интервалов. Другие способы кодирования и вычисляемые характеристики исследуются в работах [16, 17, 18].

2.2 Диагностика заболеваний по ЭКГ

Основной причиной смерти во всем мире по данным Всемирной организации здравоохранения [19] являются сердечно-сосудистые заболевания. В настоящее время разработан ряд методов машинного обучения для диагностики болезней сердца по ЭКГ [20].

Нарушение ритма. Нарушение сердечного ритма или аритмия сердца может привести к целому ряду серьезных осложнений [21]. Основным методом ее диагностики является ЭКГ (в некоторых случаях проводят электрофизиологическое исследование и картирование). Большое число исследований посвящено классификации различных видов аритмий [22]. Для экспериментов обычно используют базу данных MIT-BIH [28], содержащую 48 размеченных 30-минутных записей ЭКГ с 2 отведений. Такие методы машинного обучения, как SVM [23, 24], метод k ближайших соседей [25], нейронные сети [26], решающие деревья [27] успешно справляются с поставленной задачей (точность достигает 96% [26]).

Ишемия и инфаркт миокарда. Инфаркт миокарда — это острое проявление ишемической болезни сердца. В [29] предложен эффективный иерархический метод автоматической диагностики инфаркта миокарда (чувствительность метода составляет 92.3%, специфичность — 88.1%). Известно, что болезнь влияет на форму ST-сегмента [30], поэтому для каждого кардиоцикла кроме морфологических признаков рассчитываются коэффициенты полинома, аппроксимирующего ST-сегмент. В начале работы алгоритма все множество кардиоциклов делится на заданное число кластеров. На следующем этапе для каждой электрокардиограммы вычисляется, с какой вероятностью она принадлежит к каждому из кластеров. Вычисленные вероятности и являются признаками в финальном классификаторе. Недостатком метода является его высокая чувствительность к выбору параметров: числу кластеров и коэффициенту масштаба. Авторы статьи [31] приводят алгоритм, не требующий сложной настройки параметров и обеспечивающий более высокие значения чувствительности и специфичности на той же базе данных [32] (98.7% и 96.4% соответственно). Сигнал ЭКГ $x(t)$ представляется в виде решения дифференциального уравнения второго порядка с коэффициентами, зависящими от времени:

$$\frac{d^2x(t)}{dt^2} + b_1(t)\frac{dx(t)}{dt} + b_0(t)x(t) = 0$$

Коэффициенты оцениваются по методу наименьших квадратов, а их максимальные значения на рассматриваемом отрезке времени подаются на вход классификатору SVM. Высокое качество классификации достигается при решении задач как двухклассовой, так и многоклассовой классификации.

Артериальная гипертензия. Артериальная гипертензия (гипертония) — это стойкое повышение артериального давления, повышающее риск сердечных приступов и кровоизлияний. Авторы статьи [33] исследуют влияние болезни на частот-

ные показатели ВСР. Оказывается, что появление и прогрессирование артериальной гипертензии сопровождается снижением абсолютных значений как общей мощности спектра (TP), так и каждого из составляющих его компонентов - очень низких (VLF), низких (LF) и высоких (HF) частот. Недавние исследования [34] показывают, что ВСР-анализ позволяет предсказывать риск острых патологических состояний (инфаркта, инсульта, обмороков) с уровнем чувствительности 71.4% и специфичности 87.8%. Используемые в статье данные представляют собой 24-часовые записи ЭКГ, измеренные спустя месяц после проведения антигипертензивной терапии у 139 пациентов. К группе повышенного риска относятся 17 пациентов, испытавших инфаркт, инсульт или обморок в течение последующего года. Задача заключается в построении алгоритма, способного по записи ЭКГ предсказывать, к какой группе принадлежит пациент. Для описания ЭКГ выделяются 33 признака, основанных на анализе интервалов между R-пиками случайно выбранного 5-минутного стационарного участка ЭКГ. Авторы статьи сравнивают различные методы отбора признаков и алгоритмы классификации. Наибольшее качество классификации достигается при использовании алгоритма случайного леса без предварительного отбора признаков. Также в [34] показано, что выделение эхографических признаков дает более низкие значения чувствительности и специфичности (41.2% и 73.9% соответственно), чем выделение признаков ВСР-анализа (71.4% и 87.8%).

Синдром сонного апноэ. Апноэ во сне [35] — дыхательная пауза во время сна, определяемая как отсутствие воздушного потока в полости носа или рта, длящееся не менее 10 секунд. Состояние, при котором дыхание не останавливается, но воздушный поток существенно уменьшается, называется гипопноэ. Стандартные критерии для определения тяжести синдрома обструктивного апноэ сна — индекс апноэ/гипопноэ и индекс десатурации, отражающие среднее число всех респираторных событий за час сна. “Золотым стандартом” в диагностике синдрома сонного апноэ считается полисомнография, включающая в себя измерения ЭКГ, ЭЭГ, ЭОГ, регистрацию воздушного потока на уровне рта и носа, регистрацию дыхательных движений живота и грудной клетки и т. д. Однако, как показывают исследования, наличие синдрома можно надежно установить, используя только ЭКГ [36]. Построение линейного или квадратичного дискриминанта на признаках, получаемых из анализа временных рядов интервалов между R-зубцами ЭКГ, позволяет достичь 85% точности на тестовой выборке. Добавление признаков, выделяемых из дыхательной волны [37], увеличивает точность еще на 5%. Для диагностики синдрома сонного апноэ также используют

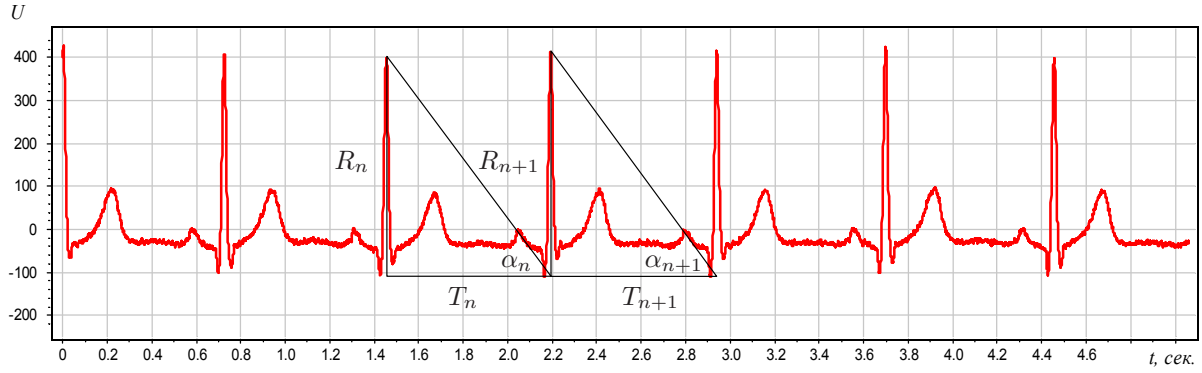


Рис. 1: Пример электрокардиограммы. Два последовательных кардиоцикла с амплитудами R_n, R_{n+1}, интервалами T_n, T_{n+1} и «фазовыми углами» α_n, α_{n+1}.

методы CART [38], SVM [39].

В отличие от известных в литературе методов, технология информационного анализа электрокардиосигналов основана на выделении коротких паттернов вариабельности, специфичных для каждого заболевания. Кроме того, учитывается не только вариабельность сердечного ритма, но и вариабельность амплитуд желудочковых QRS-комплексов.

3 Предварительная обработка ЭКГ-сигнала

Технология информационного анализа электрокардиосигналов включает три этапа их предварительной обработки: вычисление интервалов и амплитуд, дискретизацию и векторизацию [3]. Каждый этап отражает определённые предположения о том, какие свойства ЭКГ-сигнала могут иметь диагностическую ценность.

3.1 Вычисление интервалов и амплитуд

Электрокардиограмма (ЭКГ) представляет собой квазипериодический сигнал, периоды которого называются *кардиоциклами*, рис. 1.

На первом этапе обработки ЭКГ-сигнал преобразуется в последовательность пар $(R_n, T_n)_{n=1}^N$, где R_n — амплитуда, T_n — интервал n-го кардиоцикла. Также вводится арктангенс их отношения $\alpha_n = \operatorname{arctg} R_n/T_n$, как аналог фазового угла в гармонических сигналах. Последовательность $(R_n)_{n=1}^N$ называется *кардиоамплитудограммой*, $(T_n)_{n=1}^N$ — *кардиоинтервалограммой* [2]. В системе «Скринфакс», согласно применяемой методике измерений, число кардиоциклов N имеет порядок нескольких сотен,

обычно $N = 600$.

На рис. 2 показана динамика приращений амплитуд $dR_n = R_{n+1} - R_n$ и интервалов $dT_n = T_{n+1} - T_n$ в последовательных кардиоциклах ЭКГ больных и здоровых людей. Вариации амплитуд и интервалов, как правило, имеют квазипериодический характер и соответствуют дыхательной волне с периодом времени от 2 до 10 с. На них также могут оказывать влияние волна Траубе–Геринга с периодом 10–20 с, медленная волна Майера с периодом 20–300 с, сверхмедленные волны с периодом более 300 с. Таким образом, соседние кардиоциклы отражают взаимосвязи многих процессов в организме человека.

Логично предположить, что заболевания также могут привносить свои характерные паттерны в динамику приращений величин R_n, T_n, α_n. На рис. 2 хорошо видно, что динамика приращений амплитуд и интервалов кардиоциклов существенно различна у здорового человека и трёх больных различными заболеваниями. Однако остаются вопросы, насколько общезначимы эти закономерности, не являются ли они индивидуальными особенностями данных обследуемых, возможно ли выявить все такие закономерности и на их основе создать систему диагностики заболеваний.

3.2 Дискретизация

Предполагается, что диагностическую ценность имеют не столько величины амплитуд R_n, интервалов T_n и углов α_n, подверженные влиянию множества факторов, сколько знаки их приращений в последовательных кардиоциклах. Возможны только 6 сочетаний увеличений и уменьшений этих трёх величин. Эти сочетания предлагается кодировать буквами 6-символьного алфавита $\mathcal{A} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}\}$. В таблице «+» означает положительное приращение, «−» — отрицательное:

	A	B	C	D	E	F
$R_{n+1} - R_n$	+	−	+	−	+	−
$T_{n+1} - T_n$	+	−	−	+	+	−
$\alpha_{n+1} - \alpha_n$	+	+	+	−	−	−

В результате дискретизации амплитудограмма и интервалограмма преобразуются в символьную последовательность $S = (s_n)_{n=1}^{N-1}$, состоящую из символов алфавита $\mathcal{A}$ и называемую *кодограммой*, рис. 3. Каждый символ кодирует тип взаимосвязи между двумя соседними кардиоциклами. Кодограмма близка по своей сути

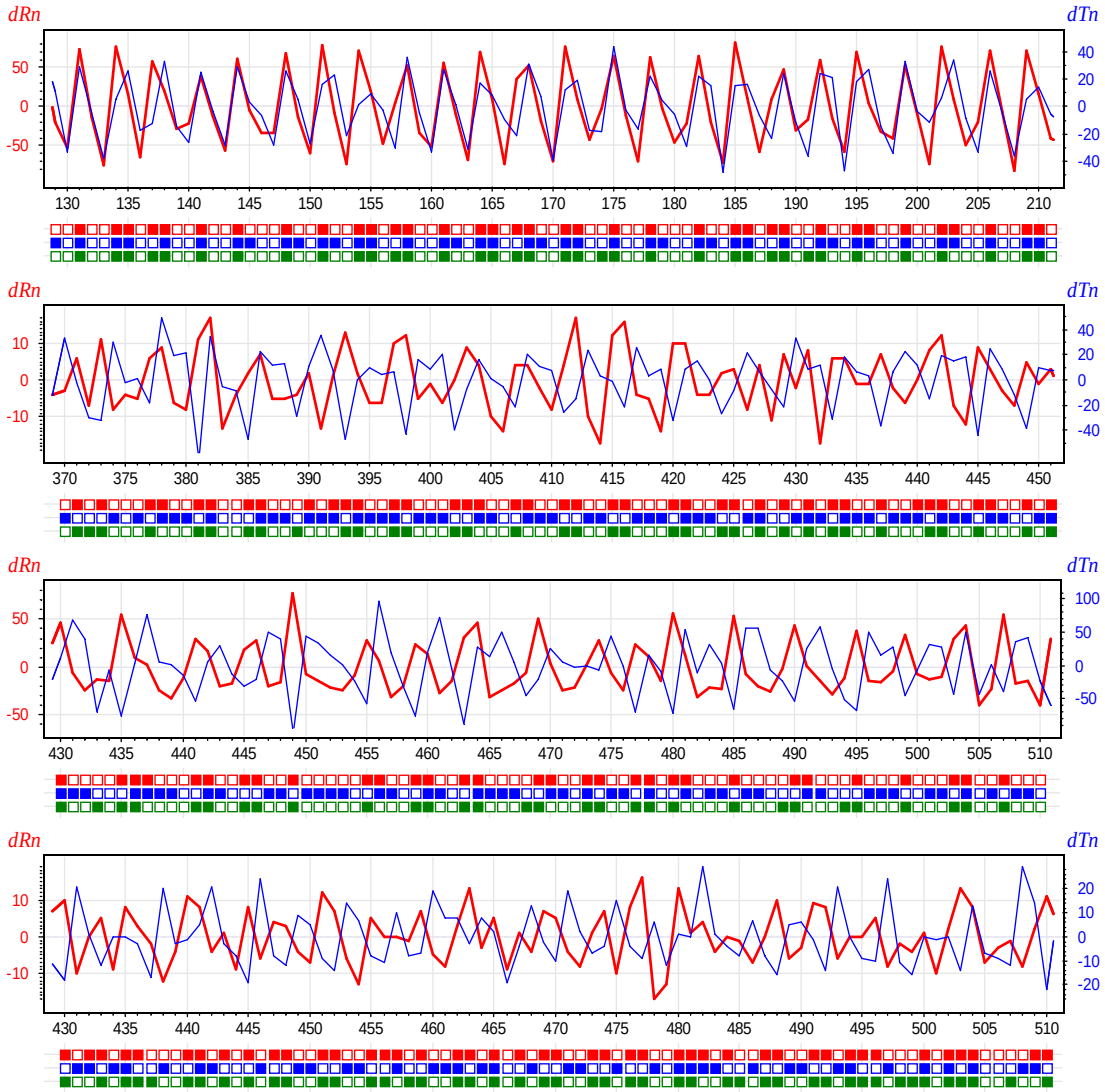


Рис. 2: Приращения амплитуд $dR_n = R_{n+1} - R_n$ (жирные красные линии) и интервалов $dT_n = T_{n+1} - T_n$ (тонкие синие линии) кардиоциклов во фрагментах ЭКГ здорового и трёх больных: язвенной болезнью, гипертонией, раком. По оси абсцисс отложены номера кардиоциклов n. Под графиками показаны знаки приращений $dR_n, dT_n, d\alpha_n$.

к тексту на естественном языке, в котором цепочки символов образуют слова, обладающие определённой семантикой. Принципиально важна способность кодограммы сохранять наиболее важную для диагностики информацию из исходного электрокардиосигнала.

Заметим, что методы анализа непрерывных сигналов путём их преобразования в символьные последовательности известны давно и находят применение в самых разных областях [40], в том числе и в анализе ЭКГ-сигналов [41]. Развиваемая в данной работе *технология информационного анализа ЭКГ-сигналов* отличается тем, что в ней учитываются вариации не только интервалов, но и амплитуд кардиоциклов.

Затем на основе их совместного анализа автоматически выявляются паттерны заболеваний и строятся диагностические правила.

Для построения диагностических правил к полученным кодограммам применяются методы анализа символьных последовательностей и машинного обучения, аналогичные тем, которые используются в вычислительной лингвистике для классификации текстов на естественном языке [42], а также в биоинформатике для классификации нуклеотидных и аминокислотных последовательностей [43].

3.3 Векторизация

Слово, образованное k последовательными буквами кодограммы $s_n, \ldots, s_{n+k-1}$, будем называть *k-граммой*, следуя терминологии вычислительной лингвистики. Множество всех возможных k-грамм $W = \mathcal{A}^k$ содержит 6^k элементов. *Частота* k-граммы $w = (w_0, \ldots, w_{k-1})$ определяется как отношение её *числа вхождений* $n_w(S)$ в кодограмму S к общему числу k-грамм в кодограмме, равному $N - k$:

$$n_w(S) = \sum_{n=1}^{N-k} \prod_{j=0}^{k-1} [s_{n+j} = w_j]; \qquad p_w(S) = \frac{n_w(S)}{N-k}.$$

В технологии информационного анализа электрокардиосигналов используются 216-мерные векторы частот триграмм, $k = 3$. Пример векторного представления кодограммы показан на рис. 4.

Дискретизация и векторизация сохраняют значимую диагностическую информацию при сокращении объёма данных в несколько тысяч раз. Существуют наборы k-грамм, совместная встречаемость которых говорит о наличии в организме *информационной сущности* или *программы* определённого заболевания. Она проявляется у человека на любой стадии заболевания, в том числе задолго до возникновения симптомов и перехода заболевания в активную фазу. Её наличие говорит о предрасположенности к заболеванию и потому может применяться с целью ранней диагностики.

3.4 Обучающая выборка

Задача построения диагностического правила по выборке больных и здоровых людей ставится как задача классификации с пересекающимися классами, поскольку у одного человека может быть много заболеваний. Обозначим через y_0 класс здоровых людей, через $y_1, \ldots, y_M$ — классы больных M заболеваниями.

Рис. 3: Пример кодограммы.

1. FFA - 42	17. EFF - 10	33. CEC - 6	49. EAC - 3
2. FAA - 33	18. DAA - 10	34. ADB - 5	50. DDA - 3
3. AFF - 32	19. ECF - 9	35. FFE - 5	51. CAC - 3
4. AAF - 30	20. FFC - 9	36. EBF - 5	52. EDF - 3
5. ADF - 18	21. FEA - 9	37. CFD - 5	53. EFB - 3
6. FCA - 18	22. DFC - 8	38. AFB - 4	54. DBA - 3
7. ACF - 17	23. ABF - 8	39. AAE - 4	55. FCC - 2
8. AAD - 15	24. AAB - 8	40. CFC - 4	56. AFC - 2
9. CFF - 14	25. FCE - 8	41. CAE - 4	57. EAA - 2
10. AEF - 13	26. AEB - 7	42. DAC - 4	58. CED - 2
11. FDA - 13	27. DFD - 7	43. DBF - 4	59. CAA - 2
12. FAE - 12	28. ACD - 6	44. BFC - 4	60. BCA - 2
13. FAC - 12	29. CDF - 6	45. CFB - 4	61. BBA - 2
14. FBA - 11	30. DFA - 6	46. AED - 3	62. DFF - 2
15. BFA - 11	31. CAF - 6	47. FFF - 3	63. BDA - 2
16. BAA - 11	32. CAD - 6	48. FBC - 3	64. DAE - 2

Рис. 4: Векторное представление $n_w(S)$ кодограммы S, приведённой на рис. 3. Показаны только 64 из 216 триграмм, имеющих число вхождений $n_w(S) \geqslant 2$.

Для каждой кодограммы из *обучающей выборки* $X = \{S_1, \ldots, S_\ell\}$ известно множество классов $Y(S_i)$. Если человек здоров, то для его кодограммы $Y(S_i) = \{y_0\}$, если же у него имеются заболевания, то y_0 не входит в $Y(S_i)$.

Важной задачей является формирование представительной и верифицированной обучающей выборки. Для всех добровольцев и больных людей, у которых осуществляли регистрацию электрокардиосигналов, проводился тщательный клинический анализ физиологических и патологических состояний с применением современных клинических, лабораторных и инструментальных методов исследования. По каждому заболеванию y_m была отобрана выборка X_m, состоящая только из тех случаев, в которых наличие специфического патоморфологического субстрата данного заболевания было надёжно установлено. Особенно тщательно была сформирована выборка X_0 здоровых людей разного пола и возраста, не имеющих существенных отклонений от состояния нормы. В экспериментах выборки больных X_m по каждому заболеванию y_m сравнивались с одной и той же выборкой здоровых X_0.

В таблице 1 перечислены заболевания и объёмы выборок до и после фильтрации аномальных записей. Аномальными считались те записи ЭКГ-сигнала, в которых более 100 из 600 кардиоциклов имели значение амплитуды R_n вне диапазона от 100 до 900 мкВ, либо значение интервала T_n вне диапазона от 400 до 1400 мс. Суммарный объём всех 19 выборок (включая АЗ) составляет 11 894 записей, что соответствует 7 216 обследованиям, из них в 5 387 случаях обследуемый имел более одного заболе-

Таблица 1: Для каждого заболевания: название, аббревиатура, код МКБ-10, объёмы выборок $|X_m|$ до и после фильтрации аномальных записей, из них число обследуемых без сопутствующих заболеваний ($|Y_m| = 1$).

анемия железодефицитная	ЖДА	D50	261	260	60
аденома простаты	ДГПЖ	N40	260	260	51
аднексит хронический	АХ	N70	276	276	33
вегетососудистая дистония	ВСД	F45.3	697	694	405
гипертоническая болезнь	ГБ	I11	1901	1894	150
асептический некроз головки бедренной кости	НГБК	M91.1	329	324	146
хронический гастрит (гастродуоденит) гиперацидный	ХГ1	K29	325	324	15
хронический гастрит (гастродуоденит) гипоацидный	ХГ2	K29	704	700	80
дискинезия желчевыводящих путей	ДЖВП	K83	717	717	10
желчнокаменная болезнь	ЖКБ	K80	278	278	40
ишемическая болезнь сердца	ИБС	I20	1269	1265	34
мочекаменная болезнь	МКБ	N20	654	654	40
миома матки	ММ	D25	784	781	190
рак общий (онкопатология различной локализации)	РО	C00-C97	531	530	9
сахарный диабет (СД1 и СД2)	СД	E10-E11	874	871	34
узловой (диффузный) зоб щитовидный железы	УЩ	D34	751	748	96
холецистит хронический	ХХ	K81.1	340	340	42
язвенная болезнь	ЯБ	K25-28	789	785	196
абсолютно здоровые	АЗ		193	193	193

вания.

4 Статистический анализ информативности триграмм

В данном разделе предлагается два перестановочных теста [44] для оценивания информативности k-грамм при $k = 3$ (триграмм). Первый тест проверяет гипотезу о наличии взаимосвязей между вариациями интервалов и амплитуд в соседних кардиоциклах. Второй тест проверяет гипотезу о наличии взаимосвязей между этими

вариациями и заболеваниями.

Взаимосвязь между частотой k-граммы и заболеваниями характеризуется *средней частотой* k-граммы w в классе y_m

$$F_w(X_m) = \frac{1}{|X_m|} \sum_{S \in X_m} p_w(S), \tag{4.1}$$

а также *встречаемостью* k-граммы w в классе y_m, равной доле кодограмм класса y_m, в которых частота k-граммы w превышает заданный порог θ или равна ему:

$$B_w(X_m, \theta) = \frac{1}{|X_m|} \sum_{S \in X_m} \big[p_w(S) \geqslant \theta\big]. \tag{4.2}$$

4.1 Тест неслучайности триграммы в последовательности кардиоциклов

Проверяется нулевая гипотеза о том, что между соседними кардиоциклами нет никакой взаимосвязи, то есть что вариации амплитуд и интервалов в соседних кардиоциклах случайны и независимы. Если эта гипотеза верна, то распределения частот триграмм не должны сильно измениться, если в последовательности $(R_n, T_n)_{n=1}^N$ случайным образом переставить местами все кардиоциклы.

Для проверки нулевой гипотезы по каждой кодограмме S_i из выборки X_m генерируется $P = 1000$ кодограмм путём случайных перестановок кардиоциклов. В результате получается P перемешанных выборок X_m^p, $p = 1, \ldots, P$. Для каждой триграммы w вычисляется минимальное и максимальное значения средней частоты F_w и встречаемости B_w:

$$F_w^{\min} = \min_{p=1,\ldots,P} F_w(X_m^p), \qquad F_w^{\max} = \max_{p=1,\ldots,P} F_w(X_m^p);$$

$$B_w^{\min} = \min_{p=1,\ldots,P} B_w(X_m^p, \theta), \qquad B_w^{\max} = \max_{p=1,\ldots,P} B_w(X_m^p, \theta).$$

Если средняя частота F_w или встречаемость B_w, вычисленные по исходной (не перемешанной) выборке, выходит за пределы найденного диапазона, то нулевая гипотеза отвергается для данной триграммы w при уровне значимости $\frac{2}{P} = 0.2\%$.

Введём также относительные величины — статистики DF_w и DB_w:

$$DF_w(X_m) = \frac{2F_w(X_m) - F_w^{\max} - F_w^{\min}}{F_w^{\max} - F_w^{\min}};$$

$$DB_w(X_m, \theta) = \frac{2B_w(X_m, \theta) - B_w^{\max} - B_w^{\min}}{B_w^{\max} - B_w^{\min}}.$$

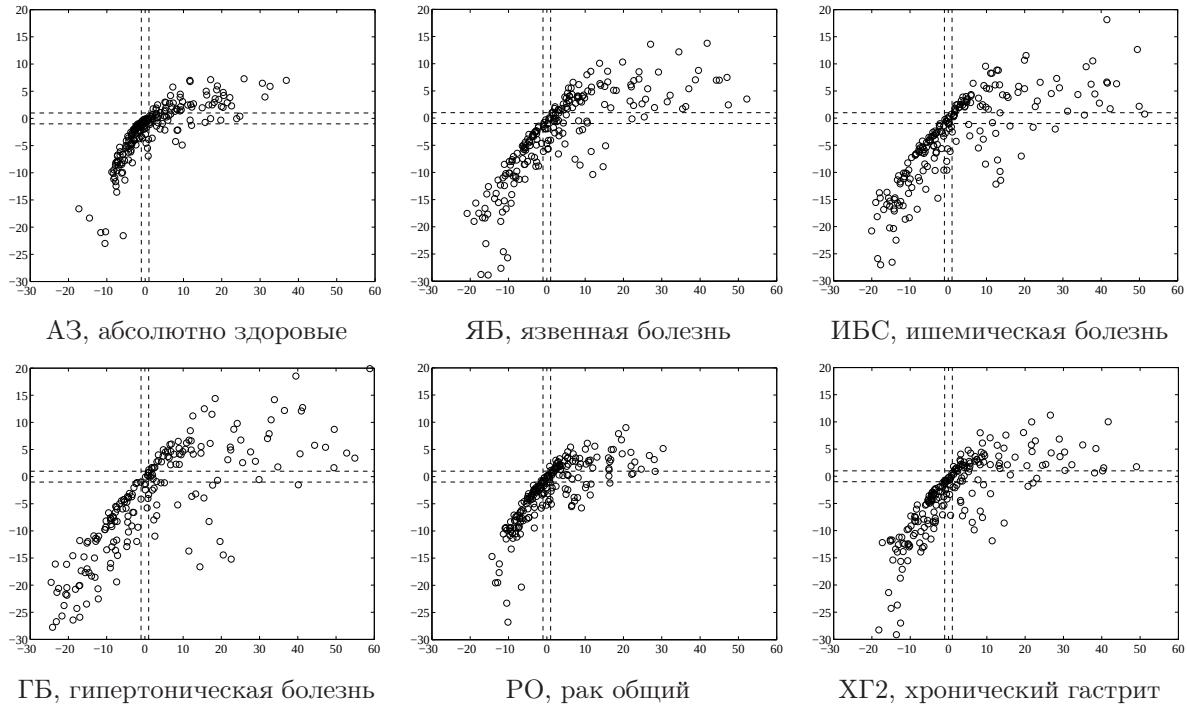


Рис. 5: Тест неслучайности триграмм в последовательности кардиоциклов. Каждая точка соответствует некоторой триграмме w. По горизонтальной оси откладываются значения DF_w, по вертикальной — DB_w при $\theta = \frac{2}{N-3}$. Выход значения DF_w или DB_w за пределы отрезка $[-1, +1]$ означает, что для триграммы w отвергается нулевая гипотеза об отсутствии взаимосвязей между составляющими её соседними кардиоциклами.

Если значение статистики DF_w или DB_w выходит за пределы отрезка $[-1, 1]$, то для триграммы w нулевая гипотеза отвергается.

Графики на рис. 5 показывают, что для пяти выборок больных и выборки здоровых данные противоречат нулевой гипотезе для большинства триграмм. Для остальных болезней графики не показаны, так как они имеют качественно тот же вид. В зависимости от заболевания нулевая гипотеза принимается для 10–38 триграмм по тесту DF_w и для 12–49 триграмм по тесту DB_w. Для остальных триграмм нулевая гипотеза отвергается, значит, составляющая их последовательность кардиоциклов является неслучайной и может нести некоторую информацию. Остаётся проверить, является ли эта информация диагностической, то есть характерна ли она для кодограмм больных данным заболеванием, или же она определяется индивидуальными особенностями обследуемых.

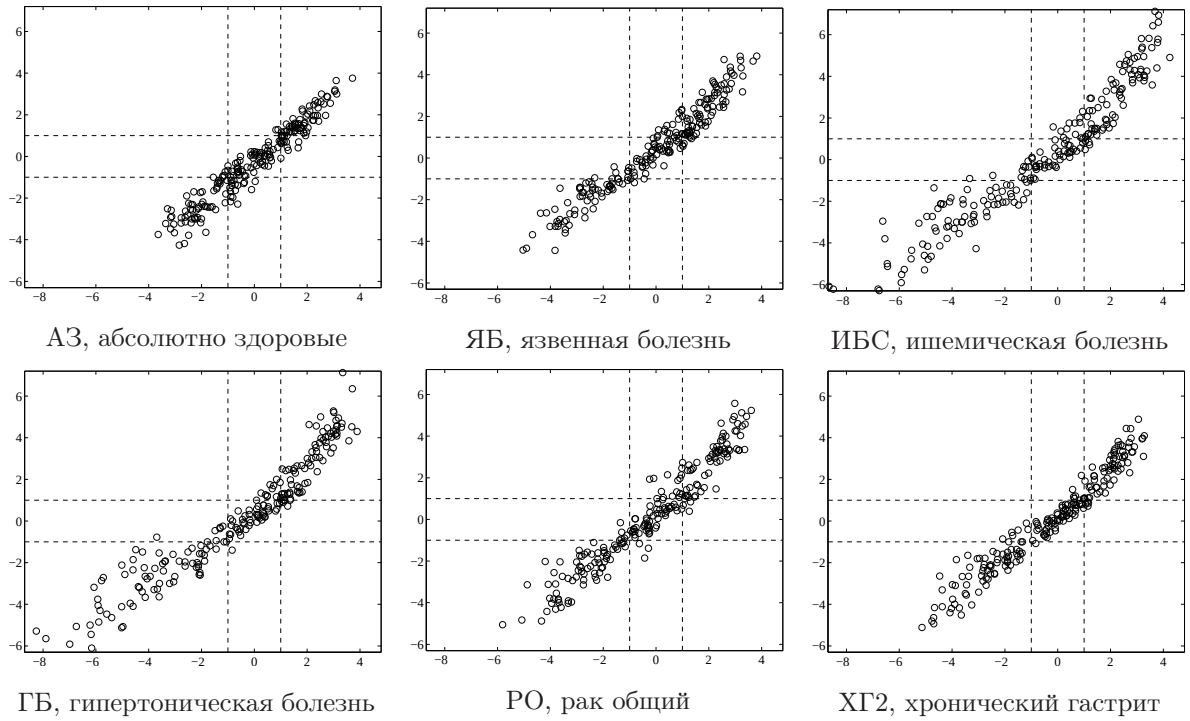


Рис. 6: Тест неслучайности триграммы относительно заболевания. Каждая точка соответствует некоторой триграмме w. По горизонтальной оси откладываются значения DF_w, по вертикальной — DB_w при $\theta = \frac{2}{N-3}$. Выход значения DF_w или DB_w за пределы отрезка $[-1, +1]$ означает, что для триграммы w отвергается нулевая гипотеза о том, что её частота не зависит от заболевания.

4.2 Тест неслучайности триграммы относительно заболевания

Проверяется нулевая гипотеза о том, что частота триграммы в выборках больных и здоровых значимо не различается, то есть что вариации амплитуд и интервалов не зависят от заболевания. Если это так, то распределение частоты триграммы не сильно изменится, если в объединённой выборке кодограмм больных и здоровых $X_m \cup X_0$ случайным образом переставить местами метки классов «больной/здоровый».

Для каждой болезни y_m генерируется $P = 1000$ перемешанных выборок X_m^p, $p = 1, \ldots, P$ путём случайных перестановок меток классов y_0 и y_m. Аналогично предыдущему тесту для каждой триграммы w вычисляется минимальное и максимальное значения средней частоты $F_w(X_m)$ и встречаемости $B_w(X_m, \theta)$ по всем P перемешанным выборкам, а также относительные величины — статистики $DF_w(X_m)$ и $DB_w(X_m, \theta)$.

Чтобы найти специфичные триграммы класса здоровых, формируется случай-

ная подвыборка больных различными заболеваниями $X_{\bar{0}}$, затем генерируется P выборок $X_0^p \cup X_{\bar{0}}^p$ с перемешанными метками классов «больной/здоровый».

Если значение статистики $DF_w(X_m)$ или $DB_w(X_m, \theta)$, вычисленное по исходной (не перемешанной) выборке X_m, выходит за пределы отрезка $[-1, 1]$, то нулевая гипотеза отвергается для данной триграммы w при уровне значимости $\frac{2}{P} = 0.2\%$.

Графики на рис. 6 показывают, что для пяти выборок больных и выборки здоровых данные противоречат нулевой гипотезе для большинства триграмм. В зависимости от заболевания нулевая гипотеза принимается для 52–96 триграмм по тесту DF_w и для 56–101 триграмм по тесту DB_w. Для остальных триграмм нулевая гипотеза отвергается, значит, они несут значимую информацию о заболевании. Каждому заболеванию соответствует свой набор триграмм, имеющих неслучайно высокие или неслучайно низкие частоты.

Графики на рис. 7 показывают расположение всех 216 триграмм $w \in W$ в осях здоровые–больные. На каждом графике из левого вертикального ряда по оси абсцисс откладывается среднее число вхождений триграммы w у здоровых $(N-3)F_w(X_0)$, по оси ординат — у больных $(N-3)F_w(X_m)$. На каждом графике из среднего вертикального ряда по осям откладываются встречаемости $B_w(X_0, \theta)$, $B_w(X_m, \theta)$ при $\theta = \frac{2}{N-3}$. На всех этих графиках вдоль диагонали располагаются триграммы, для которых нулевая гипотеза принимается. Они не являются значимыми для диагностики данного заболевания. Ступенчатыми линями показаны границы критических областей при уровне значимости 10% (ближе к диагонали) и 0.2% (дальше от диагонали). На графиках в правом вертикальном ряду показано расположение триграмм в осях встречаемостей $B_w(X_0, \theta)$, $B_w(X_m, \theta)$ для одной из 1000 случайных перестановок меток.

В критической области находятся триграммы, для которых отвергается нулевая гипотеза о независимости заболевания и частоты триграммы. Их частоты являются неслучайно высокими (точки ближе к оси ординат) или, наоборот, неслучайно низкими (точки ближе к оси абсцисс) для данного заболевания. Такие триграммы являются значимыми для диагностики данного заболевания.

Анализ графиков на рис. 7 позволяет сделать следующие выводы.

1. Для каждого заболевания имеется несколько десятков значимых триграмм.

2. Состояние нормы также характеризуется своим набором значимых триграмм.

3. Наборы значимых триграмм болезней существенно различаются, но ещё сильнее они отличаются от набора значимых триграмм класса здоровых людей.

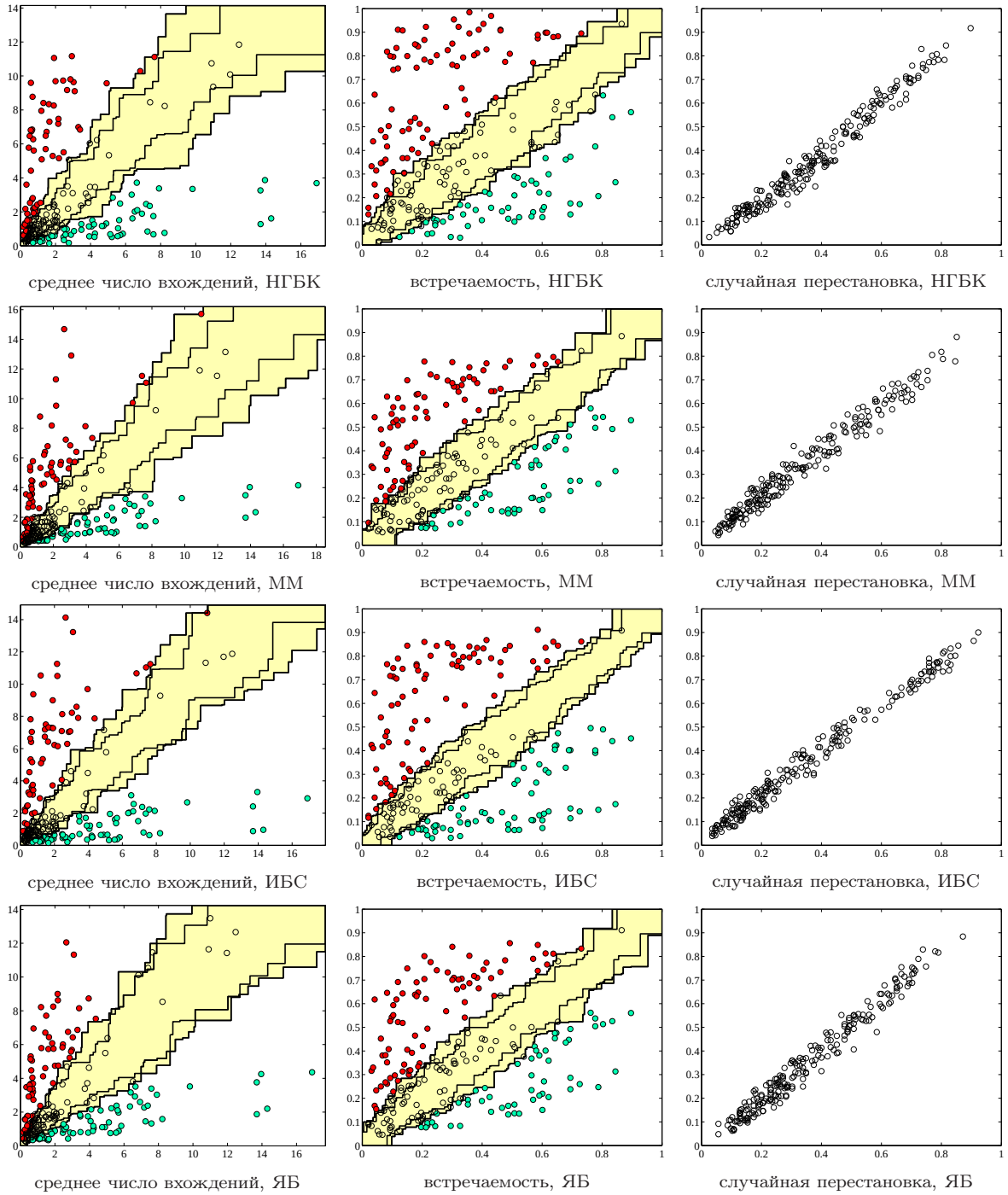


Рис. 7: Среднее число вхождений и встречаемость триграмм у здоровых и больных. Каждая точка соответствует некоторой триграмме w. По горизонтальной оси откладываются значения DF_w или DB_w у здоровых, по вертикальной оси — у больных.

4. Число незначимых триграмм с высокой средней частотой намного меньше, чем с высокой встречаемостью. Это означает, что если триграмма является значимой, то, как правило, она встречается в кодограммах данного класса много раз.

4.3 Комбинирование перестановочных тестов

Для каждого заболевания можно найти все триграммы, для которых отвергаются обе нулевые гипотезы о независимости. Чтобы найти триграммы, не только отличающие данное заболевание от класса здоровых, но и различающие заболевания между собой, предлагается модификация второго теста: выборка больных класса y_m сравнивается не с классом здоровых y_0, а с объединённой выборкой здоровых и больных всеми заболеваниями, кроме y_m.

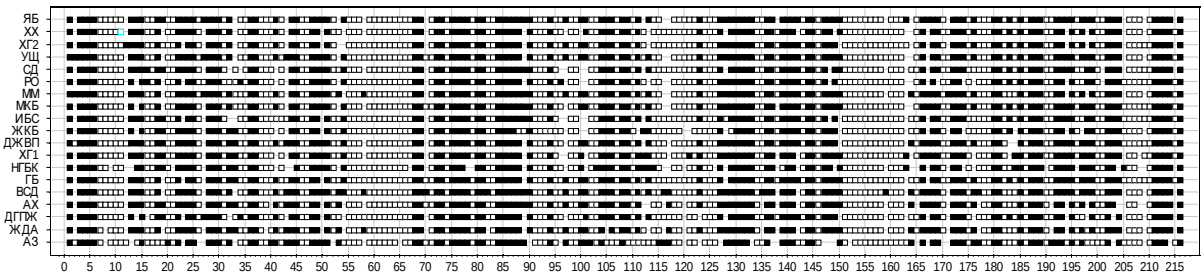
На рис. 8 приведены результаты перестановочных тестов для всех триграмм и всех заболеваний, при уровне значимости 0.2%. Согласно первому тесту 90% триграмм являются информативными, рис. 8 (а). Согласно второму тесту наборы информативных триграмм разных заболеваний заметно отличаются от триграмм класса абсолютно здоровых, но похожи между собой, рис. 8 (б). Модифицированный второй тест позволяет выявить специфичные триграммы каждого заболевания, отличающие его от всех остальных заболеваний, рис. 8 (в). Согласно этому тесту 27% триграмм являются информативными. Комбинация первого и модифицированного второго теста оставляет лишь 15% информативных триграмм, рис. 8 (д).

Таким образом, каждое заболевание характеризуется уникальным эталонным набором информативных триграмм, позволяющим отличать его как от класса здоровых, так и от других заболеваний. Эти наблюдения служат обоснованием для использования частот или встречаемостей триграмм в качестве признаков при автоматическом построении диагностических правил методами машинного обучения.

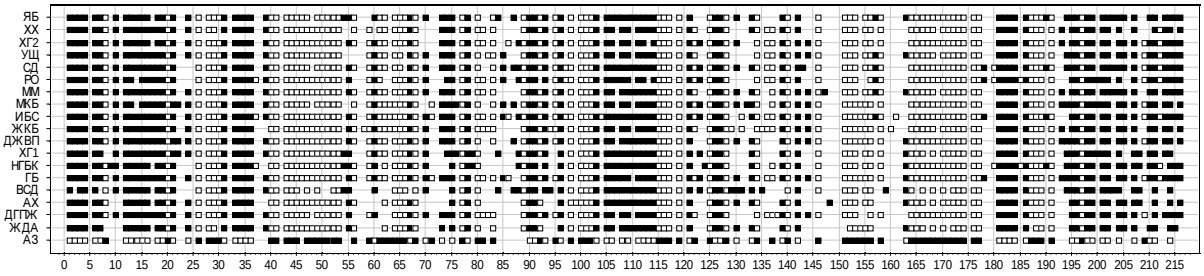
Заметим, что многие триграммы, неслучайные по первому тесту, не несут значимой информации о заболеваниях из имеющейся выборки. Однако они могут быть значимыми для других заболеваний или иных состояний организма, по которым ещё не собраны выборки данных, и которые ещё предстоит изучить.

5 Модели классификации

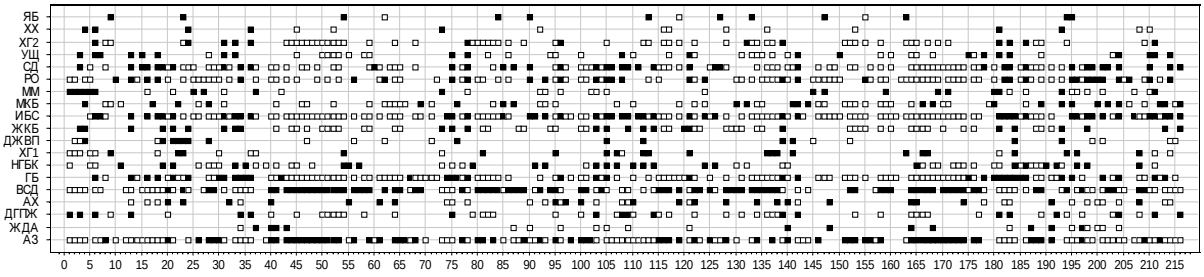
Диагностическое правило заболевания y_m — это функция $a_m(S)$, которая по кодограмме обследуемого S определяет метку класса: «0» — здоровый, «1» — больной. Классы заболеваний не являются взаимоисключающими, поэтому для каждого заболевания строится своё диагностическое правило, отличающее больных от здоровых.



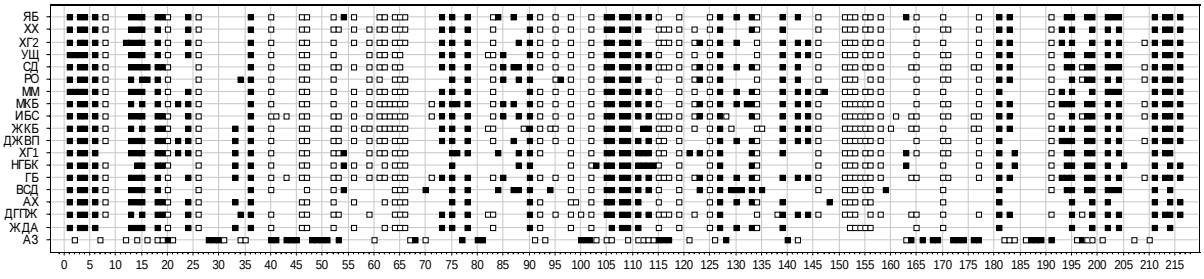
(а) тест перемешивания кардиоциклов



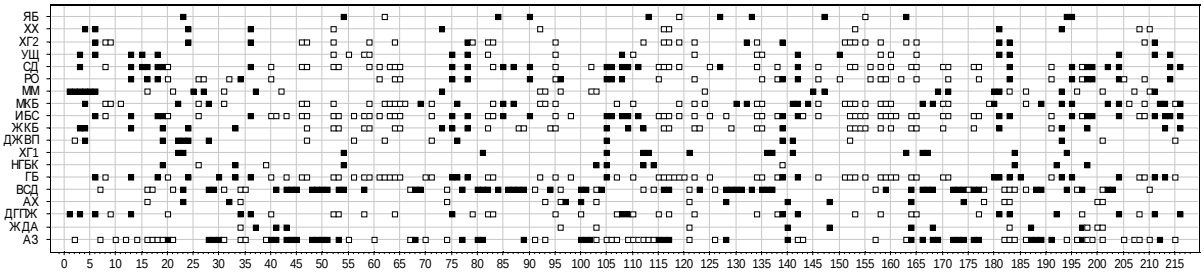
(б) тест перемешивания меток классов (болезнь против здоровых)



(в) тест перемешивания меток классов (болезнь против всех)



(г) комбинация тестов перемешивания кардиоциклов и меток классов (болезнь против здоровых)



(д) комбинация тестов перемешивания кардиоциклов и меток классов (болезнь против всех)

Рис. 8: Карты информативности триграмм, согласно тестам неслучайности и их комбинациям. По горизонтальной оси отложены все 216 триграмм, по вертикальной оси — заболевания. Чёрными точками показаны триграммы с неслучайно высокой частотой, белыми точками — с неслучайно низкой частотой.

5.1 Экспертная модель

Экспертная модель — модель классификации, использовавшаяся в ранних версиях системы «Скринфакс». Для каждого заболевания строился набор диагностических эталонов. Диагностический эталон — это набор триграмм, совместно встречающихся у определённой группы больных, но никогда совместно не встречающихся у здоровых людей. Поиск диагностических эталонов производился экспертом с помощью специально разработанной программы статистического анализа выборки кодограмм.

5.2 Линейные модели классификации

Линейными моделями классификации называются диагностические правила, в которых решение принимается по значению взвешенной суммы признаков [45, 46, 47]:

$$a_m(S) = \big[b_m(S) \geqslant \beta_m\big], \qquad b_m(S) = \sum_{w \in W} \gamma_{mw} f_w(S),$$

где $b_m(S)$ — дискриминантная функция, оценивающая степень принадлежности прецедента S классу 1; β_m — порог принятия решения; $f_w(S)$ — числовой признак, монотонно зависящий от частоты k-граммы w в кодограмме S. Задача обучения диагностического правила состоит в том, чтобы по выборкам прецедентов двух классов, здоровых X_0 и больных X_m, оптимизировать веса признаков γ_{mw}.

5.2.1 Линейные байесовские классификаторы

Согласно байесовской теории классификации, минимальным риском потерь обладает *оптимальный байесовский классификатор* вида

$$a_m(S) = \left[\ln \frac{\pi_m(S)}{\pi_0(S)} \geqslant \beta_m\right], \tag{5.1}$$

где $\pi_m(S)$ — модель плотности распределения класса y_m, порог β_m зависит от соотношения потерь от ошибок на объектах класса больных y_m и здоровых y_0.

Предположим, что k-граммы, характерные для данного заболевания, появляются в кодограмме независимо друг от друга. Тогда частоты k-грамм $p_w(S)$ являются независимыми случайными величинами. В этом случае многомерная плотность распределения $\pi_m(S)$ представляется в виде произведения одномерных плотностей,

классификатор приобретает особо простой вид и называется *наивным байесовским классификатором*. Чтобы найти одномерные плотности, предположим, что появления одной и той же k-граммы в кодограмме независимы друг от друга. Тогда число появлений $n_w(S)$ k-граммы w в кодограмме S описывается распределением Пуассона, а плотность $\pi_m(S)$ представляется в виде произведения распределений Пуассона:

$$\pi_m(S) = \prod_{w \in W} \frac{\lambda_{mw}^{n_w(S)}}{n_w(S)!} \exp(-\lambda_{mw}),$$

где λ_{mw} — параметр распределения Пуассона. Его несмещённая выборочная оценка $\lambda_{mw} = (N-k)F_w(X_m)$ совпадает со средним числом вхождений k-граммы w в кодограммы класса y_m. Подставляя эти оценки в плотности $\pi_m(S)$ и далее в формулу (5.1), получим, что оптимальный байесовский классификатор является линейным с коэффициентами γ_{mw}, которые легко вычисляются по обучающей выборке:

$$b_m(S) = \sum_{w \in W} \gamma_{mw} p_w(S), \qquad \gamma_{mw} = \ln \frac{F_w(X_m)}{F_w(X_0)}. \tag{5.2}$$

Альтернативный вариант наивного байесовского классификатора основан на предположении, что диагностическую ценность имеют не частоты k-грамм в кодограмме, а только то, какие k-граммы встречаются чаще, чем они могли бы встречаться чисто случайно. Будем полагать, что встречаемости k-грамм в каждом классе y_m — это независимые биномиальные случайные величины $f_w(S) = \big[p_w(S) \geqslant \theta\big]$ с параметрами вероятности события $\mu_{mw} = \mathsf{P}\big(p_w(S) \geqslant \theta \;\big|\; S \in X_m\big)$. Тогда

$$\pi_m(S) = \prod_{w \in W} \mu_{mw}\big[p_w(S) \geqslant \theta\big] + (1 - \mu_{mw})\big[p_w(S) < \theta\big].$$

Несмещённая выборочная оценка параметра μ_{mw} совпадает со значением встречаемости k-граммы w в выборке прецедентов класса y_m: $\mu_{mw} = B_w(X_m, \theta)$. Подставляя эти оценки в плотности $\pi_m(S)$, затем эти плотности — в формулу (5.1), снова получим линейную дискриминантную функцию, но с другими коэффициентами γ_{mw}:

$$b_m(S) = \sum_{w \in W} \gamma_{mw}\big[p_w(S) \geqslant \theta\big], \qquad \gamma_{mw} = \ln \frac{B_w(X_m, \theta)\big(1 - B_w(X_0, \theta)\big)}{B_w(X_0, \theta)\big(1 - B_w(X_m, \theta)\big)}. \tag{5.3}$$

Согласно описанным выше экспериментам, каждое заболевание y_m характеризуется своим набором информативных k-грамм. Именно эти k-граммы должны получать наибольшие по модулю веса в линейных классификаторах. Использование остальных «шумовых» k-грамм в суммах (5.2) и (5.3) может приводить к снижению информативности дискриминантной функции и падению качества классификации. Чтобы этого не происходило, предлагается ранжировать k-граммы по убыванию

некоторого критерия информативности, например, по модулю весов $|\gamma_{mw}|$, и учитывать только первые K k-грамм. Полученное множество информативных k-грамм T_m будем называть *диагностическим эталоном* заболевания y_m. Для остальных k-грамм положим $\gamma_{mw} = 0$. Число K является параметром метода и подбирается в экспериментах.

Заметим, что при $\gamma_{mw} = 1$, $w \in T_m$, величина $b_m(S)$ из (5.3) равна доле k-грамм диагностического эталона, часто встречающихся в кодограмме S. Если $b_m(S) \geqslant \beta_m$, то S относится к классу больных. Решающие правила такого вида широко используются в медицинской диагностике, когда известен набор симптомов (*синдром*) заболевания, и диагноз ставится, если у больного наблюдаются некоторые из этих симптомов. Информативные k-граммы аналогичны симптомам, а диагностические эталоны T_m — синдромам. Поэтому будем называть линейные методы классификации, основанные на выделении диагностических эталонов заболеваний, *синдромными алгоритмами*.

Рассмотрим общую модель синдромного алгоритма при различных сочетаниях типа используемых признаков, формулы весов и критерия информативности:

(F) тип признаков — вещественные частоты k-грамм $p_w(S)$:

- формула весов признаков:
  (Γ^1) $\gamma_{mw} = 1$;
  (Γ^2) $\gamma_{mw} = F_w(X_m)$;
  (Γ^3) $\gamma_{mw} = F_w(X_m) - F_w(X_0)$;
  (Γ^4) $\gamma_{mw} = \ln \tilde{F}_w(X_m) - \ln \tilde{F}_w(X_0)$;
  (Γ^5) $\gamma_{mw} = DF_w(X_m)$;
- критерий отбора K признаков с наибольшими значениями:
  (S^1) $F_w(X_m)$;
  (S^2) $F_w(X_m)\big[w \notin T_0\big]$;
  (S^3) $F_w(X_m) - F_w(X_0)$;
  (S^4) $\ln \tilde{F}_w(X_m) - \ln \tilde{F}_w(X_0)$;
  (S^5) $\big|\ln \tilde{F}_w(X_m) - \ln \tilde{F}_w(X_0)\big|$;
  (S^6) $DF_w(X_m)$;
  (S^7) $\big|DF_w(X_m)\big|$;

(B) тип признаков — бинарные встречаемости $\big[p_w(S) \geqslant \theta\big]$ с параметром θ:

- формула весов признаков:
  (Γ^1) $\gamma_{mw} = 1$;
  (Γ^2) $\gamma_{mw} = B_w(X_m, \theta)$;
  (Γ^3) $\gamma_{mw} = B_w(X_m, \theta) - B_w(X_0, \theta)$;
  (Γ^4) $\gamma_{mw} = \ln \tilde{B}_w(X_m, \theta) - \ln \tilde{B}_w(X_0, \theta)$;
  (Γ^5) $\gamma_{mw} = \ln \tilde{B}_w(X_m, \theta)(1 - \tilde{B}_w(X_0, \theta)) - \ln \tilde{B}_w(X_0, \theta)(1 - \tilde{B}_w(X_m, \theta))$;
  (Γ^6) $\gamma_{mw} = DB_w(X_m, \theta)$;
- критерий отбора K признаков с наибольшими значениями:
  (S^1) $B_w(X_m, \theta)$;
  (S^2) $B_w(X_m, \theta)\big[w \notin T_0\big]$;
  (S^3) $B_w(X_m, \theta) - B_w(X_0, \theta)$;
  (S^4) $\ln \tilde{B}_w(X_m, \theta) - \ln \tilde{B}_w(X_0, \theta)$;
  (S^5) $\big|\ln \tilde{B}_w(X_m, \theta) - \ln \tilde{B}_w(X_0, \theta)\big|$;
  (S^6) $\ln \tilde{B}_w(X_m, \theta)(1 - \tilde{B}_w(X_0, \theta)) - \ln \tilde{B}_w(X_0, \theta)(1 - \tilde{B}_w(X_m, \theta))$;
  (S^7) $\big|\ln \tilde{B}_w(X_m, \theta)(1 - \tilde{B}_w(X_0, \theta)) - \ln \tilde{B}_w(X_0, \theta)(1 - \tilde{B}_w(X_m, \theta))\big|$;
  (S^8) $DB_w(X_m, \theta)$;
  (S^9) $\big|DB_w(X_m, \theta)\big|$.

Под логарифмом вместо средних частот и встречаемостей используются сглаженные байесовские оценки параметров λ_{mw} и μ_{mw}, которые всегда больше нуля:

$$\tilde{F}_w(X_m) = \frac{1}{|X_m| + 1}\left(\sum_{S \in X_m} p_w(S) + \frac{2}{N - k}\right);$$
$$\tilde{B}_w(X_m, \theta) = \frac{1}{|X_m| + 2}\left(\sum_{S \in X_m} \big[p_w(S) \geqslant \theta\big] + 1\right).$$

Таким образом, имеется семейство из $5 \cdot 7 + 6 \cdot 9 = 89$ моделей, с параметрами θ, K. Из этих моделей предлагается выбирать лучшую для каждого заболевания.

Введённые обозначения позволяют записывать модели с помощью компактных формул. Например, две введённые выше разновидности байесовского классификатора с отбором признаков по максимальным модулям весов являются частными случаями синдромного алгоритма $F\Gamma^4 S^5$ и $B\Gamma^5 S^7$.

5.2.2 Логистическая регрессия

Наивный байесовский классификатор основан на жёстком предположении о независимости признаков. Поэтому в качестве альтернативы будем определять

коэффициенты в линейном классификаторе (5.3) методом логистической регрессии с определением оптимальной размерности информативного подпространства признаков.

По-прежнему будем рассматривать два типа признаков: вещественные частоты триграмм $p_w(S)$ и бинарные встречаемости $\left[p_w(S) \geqslant \theta\right]$ с параметром θ.

Рассмотрим следующие методы отбора признаков или понижения размерности.

- Отбор K признаков с наибольшими значениями критерия информативности:
  1) $B_w(X_m, \theta)$ — встречаемость k-граммы в классе больных;
  2) $F_w(X_m)$ — частота k-граммы в классе больных;
  3) $DB_w(X_m, \theta)$ — статистика теста неслучайности относительно заболевания;
  4) $DF_w(X_m)$ — статистика теста неслучайности относительно заболевания.

- Логистическая регрессия с L_1-регуляризацией — метод LASSO [47].

- Логистическая регрессия на первых K главных компонентах [47], которые являются линейными комбинациями исходных признаков.

В методе главных компонент используются все исходные признаки (k-граммы) и определяется оптимальная размерность линейного подпространства, в котором классы наилучшим образом разделяются линейной поверхностью. Параметры θ, K подбираются в экспериментах, отдельно для каждого заболевания.

5.3 Случайный лес

Случайный лес [48] в настоящее время считается одним из самых эффективных алгоритмов машинного обучения. Он представляет собой ансамбль решающих деревьев, каждое из которых строится по случайным подвыборкам, полученным в результате сэмплирований с возвращениями объектов обучающей выборки (бэггинг [49]). Кроме того, при создании очередного узла каждого дерева выбор признака, на основе которого происходит разбиение, производится не из всего множества признаков, а из их случайного подмножества (метод случайных подпространств [50]). Классификация объектов проводится путем простого голосования: каждое решающее дерево относит объект к одному из классов, решение принимается на основании большинства голосов. Для оценки качества построенного классификатора удобно ввести дискриминантную функцию, оценивающую степень принадлежности объекта к классу, например, равную доле деревьев, голосующих за этот класс.

Пусть C_m — множество деревьев в ансамбле, построенном по обучающей выборке здоровых и больных заболеванием y_m; $g_{c_m}(S)$ — класс, к которому дерево $c_m \in C_m$ относит объект S. Тогда решающее правило представляется в виде:

$$a_m(S) = \big[b_m(S) \geqslant \beta_m\big], \qquad b_m(S) = \frac{1}{|C_m|} \sum_{c_m \in C_m} \big[g_{c_m}(S) = 1\big],$$

где $b_m(S)$ — доля деревьев, относящих прецедент S к классу 1 (дискриминантная функция); β_m — порог принятия решения (равен 1/2 при простом голосовании деревьев).

6 Эксперименты

В данном разделе приводятся результаты обучения классификаторов с помощью синдромного алгоритма, логистической регрессии и случайного леса при использовании 2-, 3- и 4-грамм. Для настройки параметров и выбора лучшей версии каждого из алгоритмов используется стандартная методика 10-кратной кросс-валидации; итоговое качество диагностики оценивается по отложенной выборке.

6.1 Оценивание качества диагностики

Для измерения качества классификации в медицинской диагностике принято использовать меры чувствительности и специфичности.

Чувствительность — это доля больных, для которых диагностическое правило верно диагностирует наличие болезни:

$$\text{чувствительность} = \frac{1}{|X_m|} \sum_{S \in X_m} \big[a_m(S) = 1\big].$$

Специфичность — это доля здоровых, для которых диагностическое правило верно диагностирует отсутствие болезни:

$$\text{специфичность} = \frac{1}{|X_0|} \sum_{S \in X_0} \big[a_m(S) = 0\big].$$

Варьируя порог β_m, можно подбирать компромисс между чувствительностью и специфичностью. Поскольку выбор такого компромисса зависит от индивидуальных предпочтений в каждом конкретном случае, используется также мера качества диагностики AUC, не зависящая от выбора порога β_m, и характеризующая только

качество дискриминантной функции $b_m(S)$. AUC (Area Under Curve) — это площадь под кривой, отображающей зависимость чувствительности от специфичности. Также AUC можно определять как долю правильно упорядоченных пар прецедентов:

$$\text{AUC} = \frac{1}{|X_0| \cdot |X_m|} \sum_{S \in X_0} \sum_{S' \in X_m} \big[b_m(S) < b_m(S') \big].$$

Для оценивания качества диагностики чувствительность и специфичность нельзя вычислять по той же выборке, по которой происходило обучение правила, так как такие оценки обязательно получатся оптимистично завышенными. Этот эффект называется *переобучением* или переподгонкой (overfitting). Чтобы его избежать, выборка случайным образом делится на 2 части: обучающую (80% объектов) и контрольную (20% объектов). По первой части выборки проводится настройка параметров модели с помощью методики *10-блочной кросс-валидации* [47], по второй — оценивание итогового качества диагностики. На стадии обучения выборка разбивается случайным образом на 10 блоков равного объёма, каждый блок по очереди становится тестовой выборкой, объединение остальных 9 блоков образует обучающую выборку. Обучение производится 10 раз, в результате каждый объект исходной выборки ровно один раз классифицируется как тестовый. По этим классификациям для каждого набора параметров вычисляются оценки AUC; значения параметров, при которых AUC достигает максимума, считаются оптимальными для данной модели.

Вычислительные эксперименты проводятся на данных по 18 заболеваниям, перечисленным в таблице 1.

6.2 Результаты работы синдромного алгоритма

- На рис. 9 слева показаны значения AUC на тестовых выборках кросс-валидации для лучших версий синдромного алгоритма при использовании вещественных и бинарных признаков. Для всех болезней бинаризация признаков повышает качество классификации. На рис. 9 справа показаны значения AUC при $k = 2, 3, 4$. В целом, увеличение k ведет к увеличению качества классификации.

- На рис. 10–12 для каждой болезни версии синдромного алгоритма при $k = 2, 3, 4$ выделены в соответствии с их рангами, т. е. в соответствии с позициями в отсортированном по убыванию качества классификации списке. Выбор формулы весов слабее влияет на качество классификации, чем выбор критерия

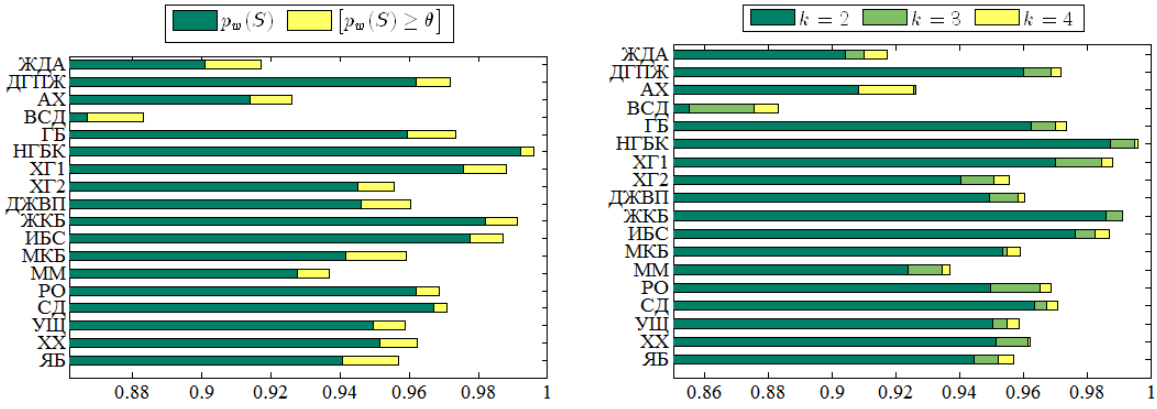


Рис. 9: Значения AUC при использовании различных типов признаков (слева) и значений k (справа), синдромный алгоритм.

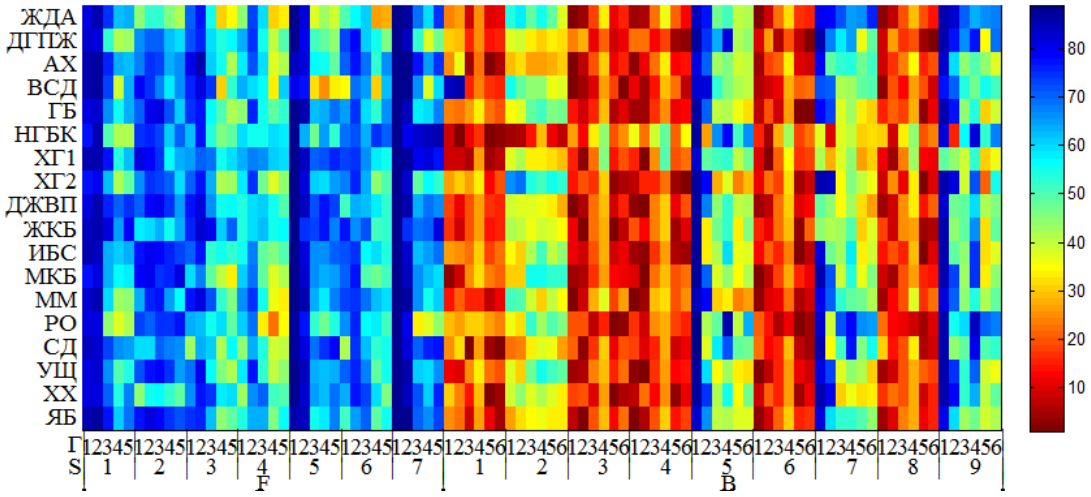


Рис. 10: Ранги версий синдромного алгоритма при $k = 2$

информативности. Критерии информативности S^1, S^3, S^4, S^6 и S^8 при бинаризации признаков для всех k показывают наилучшие результаты. В дальнейшем используется версия алгоритма $B\Gamma^5S^1$ с отбором признаков по критерию $B_w(X_m, \theta)$ и весами, равными:

$$\gamma_{mw} = \ln \tilde{B}_w(X_m, \theta)(1 - \tilde{B}_w(X_0, \theta)) - \ln \tilde{B}_w(X_0, \theta)(1 - \tilde{B}_w(X_m, \theta)).$$

На рис. 13–14 для каждой болезни на левом графике показана зависимость AUC от параметра K для синдромного алгоритма $B\Gamma^5S^1$ при $k = 4$ и $\theta = 1/(N - k)$.

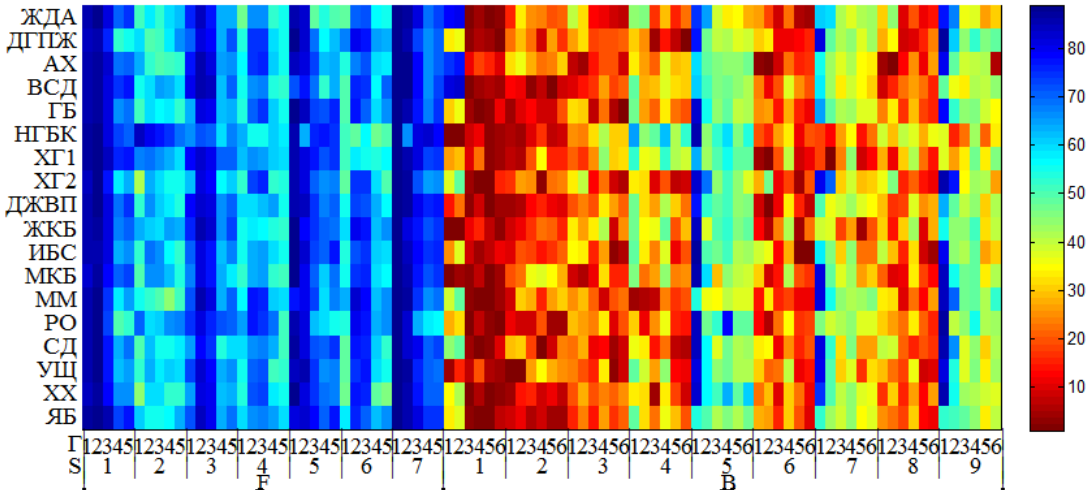


Рис. 11: Ранги версий синдромного алгоритма при $k = 3$

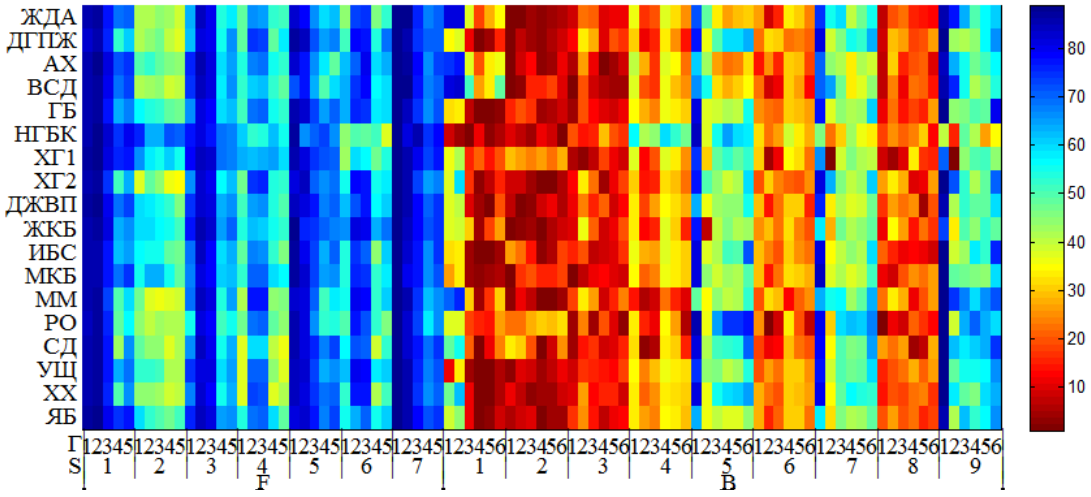
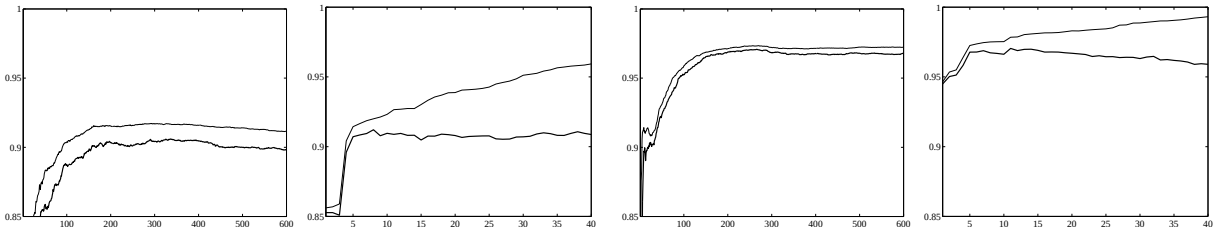


Рис. 12: Ранги версий синдромного алгоритма при $k = 4$

6.3 Результаты работы логистической регрессии

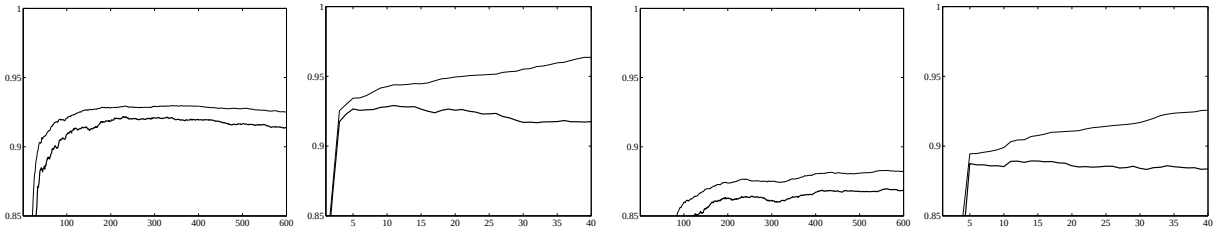
- На рис. 15 слева показаны результаты классификации логистической регрессии при использовании вещественных и бинаризованных признаков. Значения AUC для разных $k = 2, 3, 4$ приведены на рис. 15 справа. Так же, как и для синдромного алгоритма, бинаризация признаков и использование 4-грамм для большинства болезней ведет к увеличению AUC.

- На рис. 16 приведены результаты работы алгоритма при отборе бинаризованных признаков с наибольшими значениями критериев информативности $B_w(X_m, \theta)$, $F_w(X_m)$, $DB_w(X_m, \theta)$, $DF_w(X_m)$, при использовании L_1-



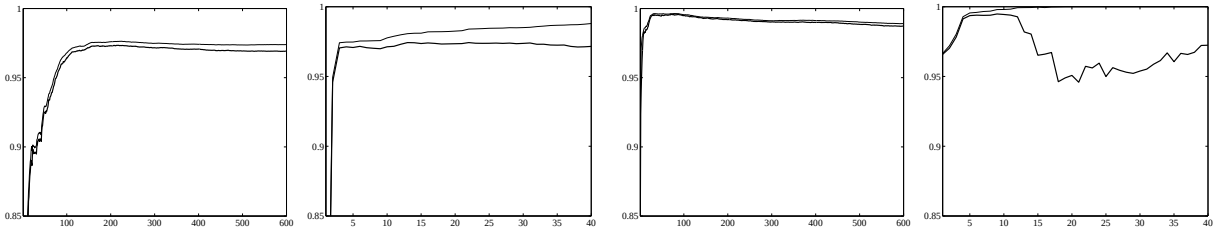
ЖДА, анемия железодефицитная

ДГПЖ, аденома простаты



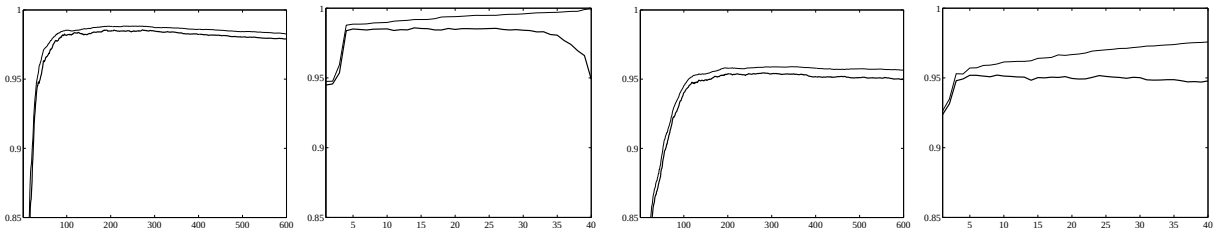
АХ, аднексит хронический

ВСД, вегетососудистая дистония



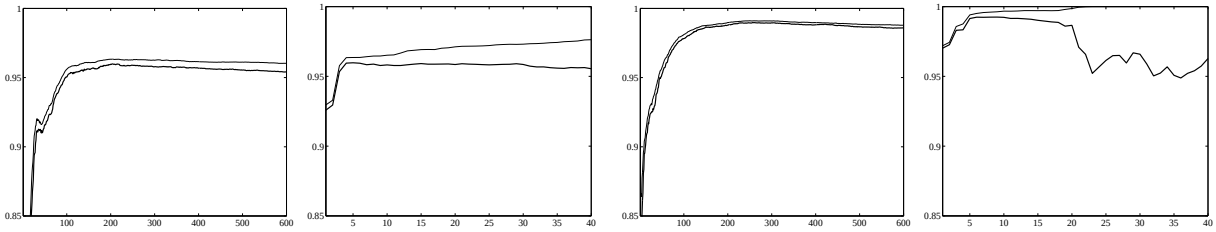
ГБ, гипертоническая болезнь

НГБК, некроз головки бедренной кости



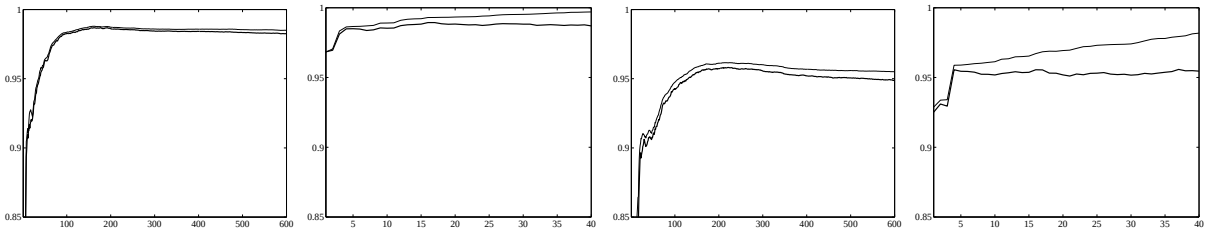
ХГ1, хронический гастрит

ХГ2, хронический гастрит



ДЖВП, дискинезия желчевыводящих путей

ЖКБ, желчнокаменная болезнь



ИБС, ишемическая болезнь

МКБ, мочекаменная болезнь

Рис. 13: Зависимости AUC на обучении (тонкие линии) и тесте (жирные линии) от числа отобранных 4-грамм K для синдромного алгоритма (левый график для каждой болезни) или от числа главных компонент K для логистической регрессии (правый график).

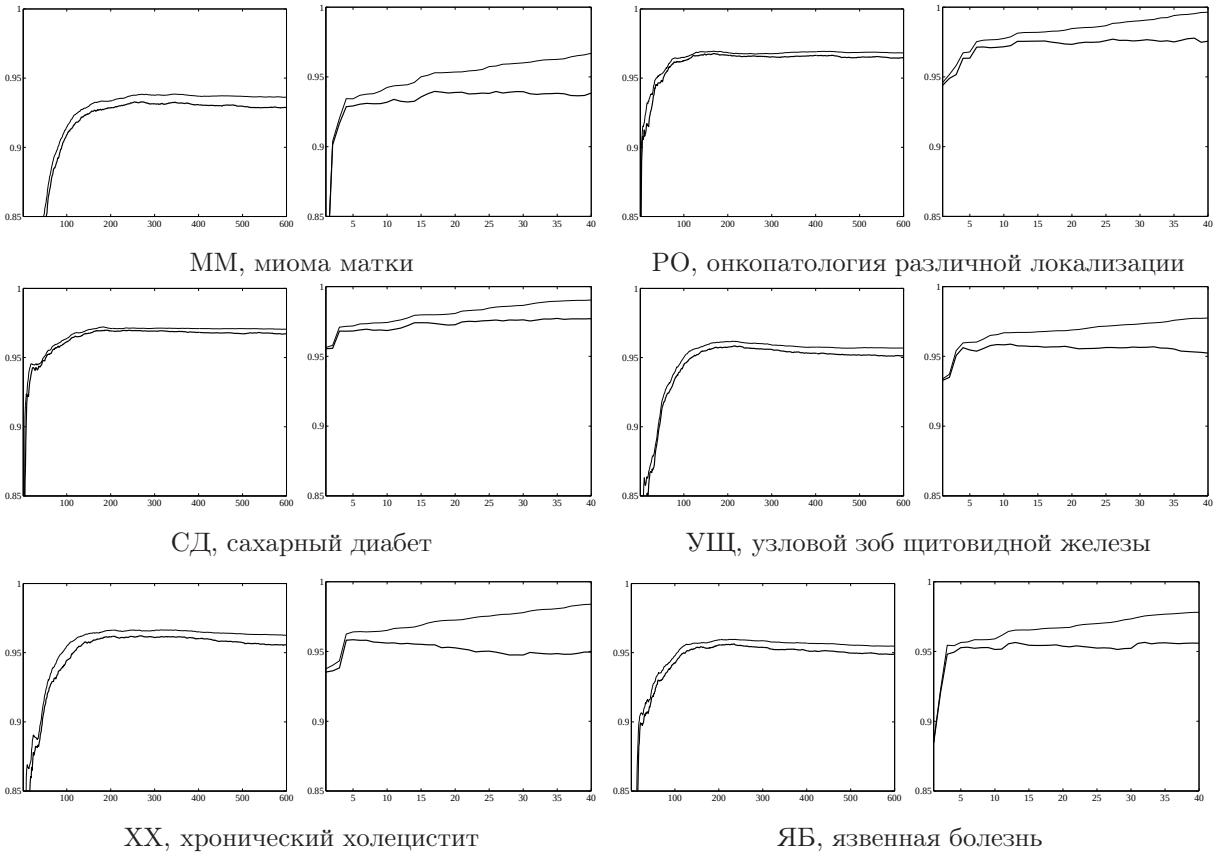


Рис. 14: Зависимости *AUC* на обучении (тонкие линии) и тесте (жирные линии) от числа отобранных 4-грамм K для синдромного алгоритма (левый график для каждой болезни) или от числа главных компонент K для логистической регрессии (правый график).

регуляризации и метода главных компонент (PCA). Для всех $k = 2, 3, 4$ наибольший AUC на контроле достигается при использовании метода главных компонент; среди остальных методов отбора признаков, наилучшим оказывается метод L_1-регуляризации.

На рис. 13–14 для каждой болезни на правом графике показана зависимость AUC от числа главных компонент K при бинаризации признаков - частот 4-грамм. Минимальное достаточное число главных компонент K, при котором AUC на тестовых данных отличается от максимального AUC не более чем на 0.5%, принимает значения от 4 до 8 для всех болезней. Существование информативных подпространств очень низкой размерности принято считать косвенным свидетельством адекватности модели классификации.

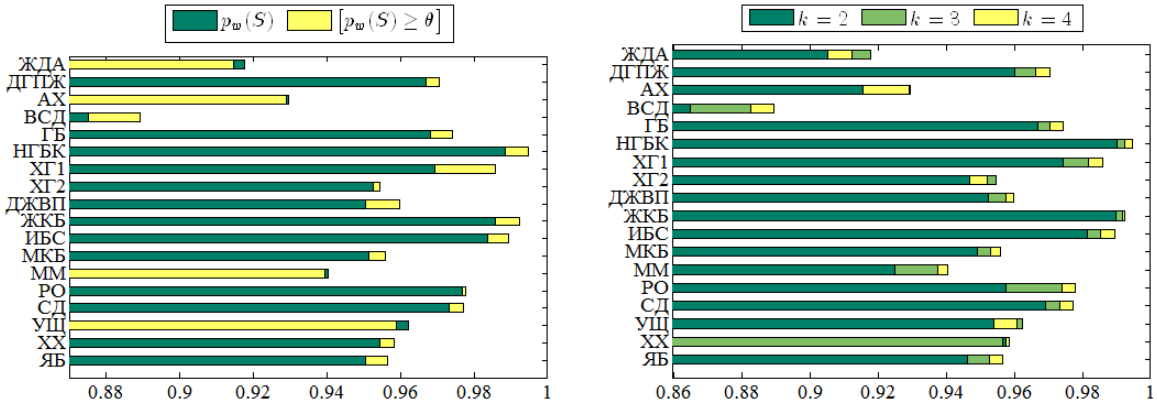


Рис. 15: Значения AUC при использовании различных типов признаков (слева) и значений k (справа), логистическая регрессия.

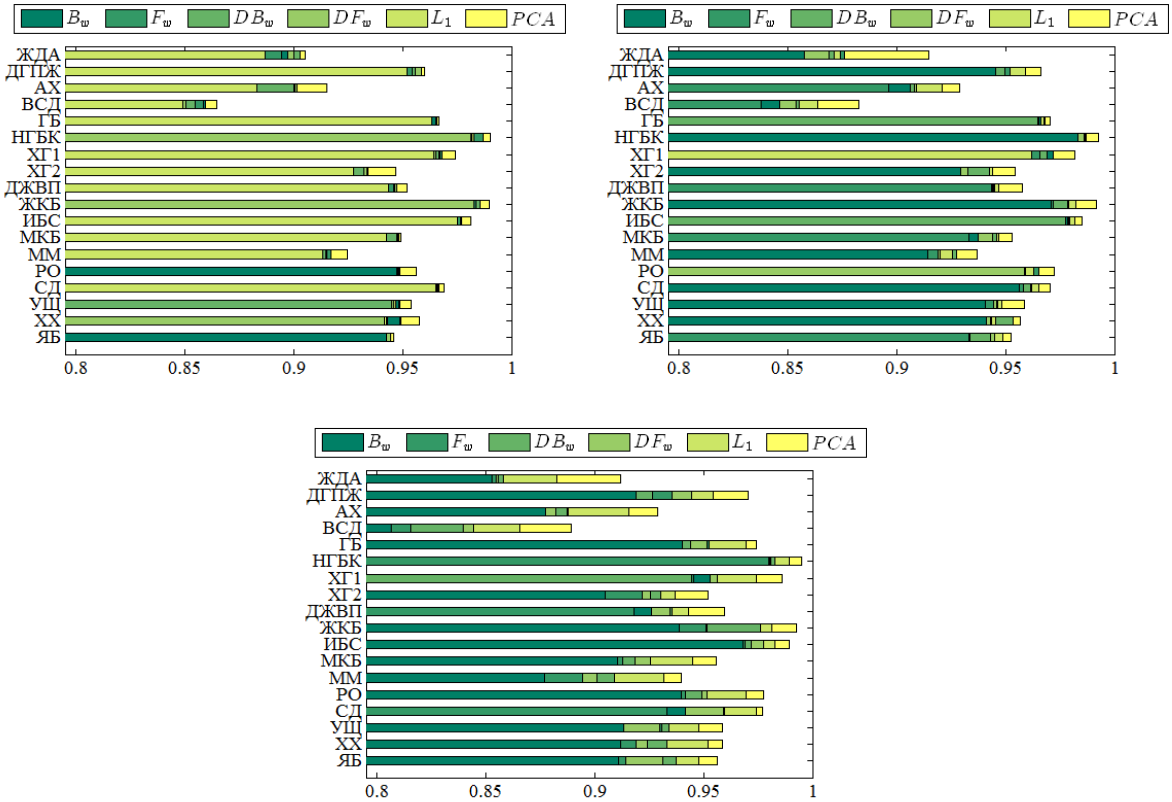


Рис. 16: Значения AUC при использовании различных методов отбора признаков при $k = 2, 3, 4$, логистическая регрессия.

6.4 Результаты работы случайного леса

На рис. 18 изображена зависимость AUC на тестовых выборках кросс-валидации от числа деревьев. В качестве признаков рассматривались частоты 3-грамм (для

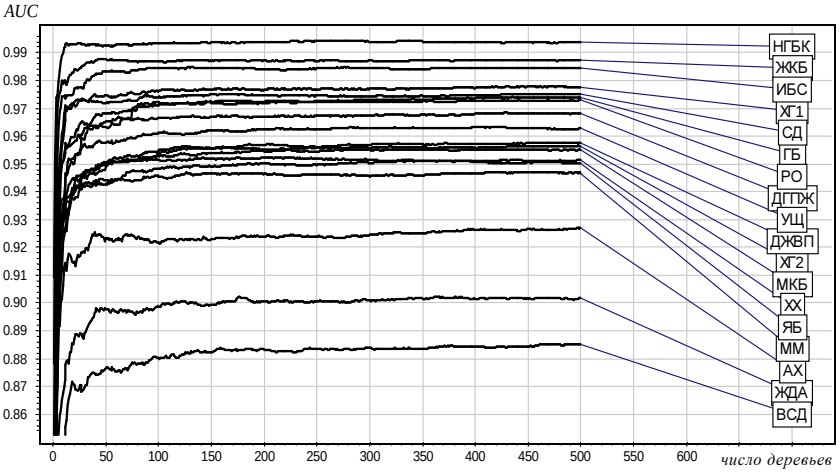


Рис. 17: Зависимость AUC от числа деревьев

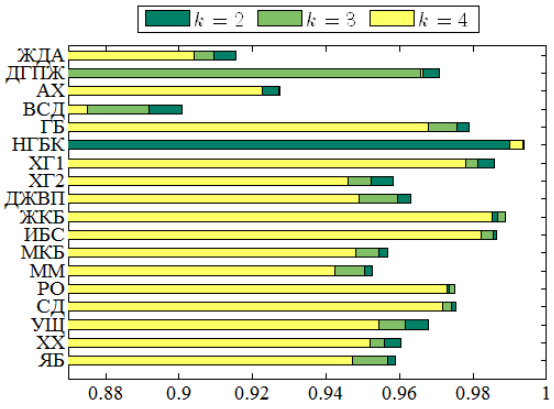


Рис. 18: Значения AUC при $k = 2, 3, 4$, случайный лес.

2- и 4-грамм зависимость имеет аналогичный вид). Т. к. качество классификации слабо зависит от числа деревьев, то дальнейшие эксперименты проведены при числе деревьев 300. На рис. 18 показаны значения AUC при построении случайного леса на признаках - частотах 2-, 3- или 4-грамм. В отличие от синдромного алгоритма и логистической регрессии, наибольшие значения AUC достигаются при использовании 2- или 3-грамм.

6.5 Результаты сравнения методов

В таблице 2 приведены значения AUC на тестовых выборках кросс-валидации для следующих методов классификации: НМ — экспертная модель, SA — версия $B\Gamma^5S^1$ синдромного алгоритма, LR — логистическая регрессия на главных ком-

понентах, построенная по бинаризованным признакам, RF — случайный лес. Через двоеточие для каждого алгоритма указано число k, определяющее тип используемых признаков (2-, 3- или 4-граммы).

Экспертная модель (HM) имеет относительно низкое качество диагностики, остальные 3 модели (SA, LR, RF) дают сопоставимое качество классификации: различия AUC не превышают 1% для всех классов заболеваний, кроме ВСД и ММ.

В синдромном алгоритме каждое заболевание характеризуется набором наиболее информативных k-грамм, образующих уникальный диагностический эталон данного заболевания. Синдромный алгоритм практически не подвержен переобучению и очень устойчив. Варьирование мощности K диагностического эталона в окрестности оптимального значения слабо влияет на качество диагностики. Оптимальное значение параметра K можно подбирать по обучающей выборке, вообще не выделяя тестовую выборку и используя весь объём данных для обучения.

Логистическая регрессия более подвержена переобучению. Кривые AUC на обучении и на тесте заметно расходятся при превышении оптимального числа главных компонент K. Метод чувствителен к отклонениям K от оптимального значения, и оптимизация K возможна только по тестовой выборке.

Преимущество методов HM и SA в том, что они строят модели простой структуры, основанные на выделении одного или нескольких диагностических эталонов для каждого заболевания. Методы LR и RF строят плохо интерпретируемые модели (т. н. «чёрные ящики»), в которых трудно выделять кодовые образы заболеваний.

ROC-кривые показывают зависимость чувствительности от специфичности, рис. 19. По оси абсцисс ROC-кривой откладывается доля ошибочных классификаций среди здоровых ($1 -$ специфичность), по оси ординат — доля верных классификаций среди больных (чувствительность). Чем левее и выше проходит график ROC-кривой, тем лучше. Каждая точка ROC-кривой представляет компромисс между чувствительностью и специфичностью.

Все кривые построены по контрольной выборке. Параметры алгоритмов настроены по тестовым выборкам кросс-валидации. В Таблице 3 сведены оценки чувствительности, специфичности и площади AUC на контрольной выборке. Все используемые алгоритмы классификации позволяют выбирать баланс чувствительности и специфичности с помощью порога β_m. Поэтому приводятся результаты двух измерений качества. В первом измеряется специфичность при фиксированной чувствительности 95%. Во втором чувствительность и специфичность имеют близкие значения, для

	НМ:3	SA:2	SA:3	SA:4	LR:2	LR:3	LR:4	RF:2	RF:3	RF:4
НГБК	98,42	98,72	99,44	**99,59**	99,02	99,25	99,48	98,99	99,35	99,34
ЖКБ	97,19	98,45	99,03	98,98	98,99	99,19	**99,25**	98,67	98,87	98,53
ИБС	95,04	97,31	98,23	98,69	98,15	98,54	**98,94**	98,64	98,56	98,21
ХГ1	93,85	96,98	98,40	98,57	97,43	98,18	**98,59**	98,58	98,12	97,80
ГБ	89,80	95,39	96,94	97,34	96,68	97,05	97,43	**97,90**	97,57	96,78
РО	—	94,42	96,49	96,78	95,64	97,22	**97,78**	97,35	97,49	97,29
СД	93,54	96,25	96,63	96,99	96,92	97,06	**97,72**	97,54	97,42	97,16
ДГПЖ	94,50	95,75	96,79	**97,08**	96,01	96,61	97,04	97,08	96,58	96,58
УЩ	88,95	94,40	95,36	95,84	95,38	95,88	95,87	**96,78**	96,15	95,43
ДЖВП	88,40	94,59	95,82	96,01	95,23	95,75	95,97	**96,32**	95,94	94,90
ХХ	90,61	95,12	96,13	**96,23**	95,76	95,66	95,84	96,02	95,58	95,20
ЯБ	91,54	94,06	95,18	95,64	94,60	95,25	95,65	**95,88**	95,68	94,72
ХГ2	80,18	93,69	94,98	95,45	94,69	95,45	95,20	**95,82**	95,22	94,60
МКБ	75,64	94,43	95,46	**95,81**	94,91	95,28	95,58	95,68	95,42	94,81
ММ	79,45	91,81	93,47	93,29	92,47	93,71	93,96	**95,25**	95,04	94,25
АХ	—	90,84	92,31	92,18	91,53	**92,91**	92,91	92,70	92,73	92,26
ЖДА	77,89	89,40	91,01	90,61	90,51	91,46	91,21	**91,55**	90,95	90,42
ВСД	72,63	84,87	87,41	86,97	86,47	88,25	88,94	**90,08**	89,17	87,51

Таблица 2: Значения AUC на тестовых выборках кросс-валидации при использовании разных методов классификации и типов признаков

них в таблице приводится общее среднее значение.

6.6 Выбор способа кодирования

Анализ ВСР использует только данные о вариабельности RR-интервалов. Поэтому важно показать, что учёт особенностей совместной вариабельности интервалов, амплитуд и их отношений даёт значимый прирост качества диагностики, причём не только для заболеваний сердечно-сосудистой системы (ИБС, ГБ), но и для широкого спектра заболеваний внутренних органов.

На рис. 20 показаны результаты работы случайного леса из 100 деревьев на тестовый выборках кросс-валидации при выборе различных способов кодирования: учитывающих только приращения амплитуд (2R), только приращения интерва-

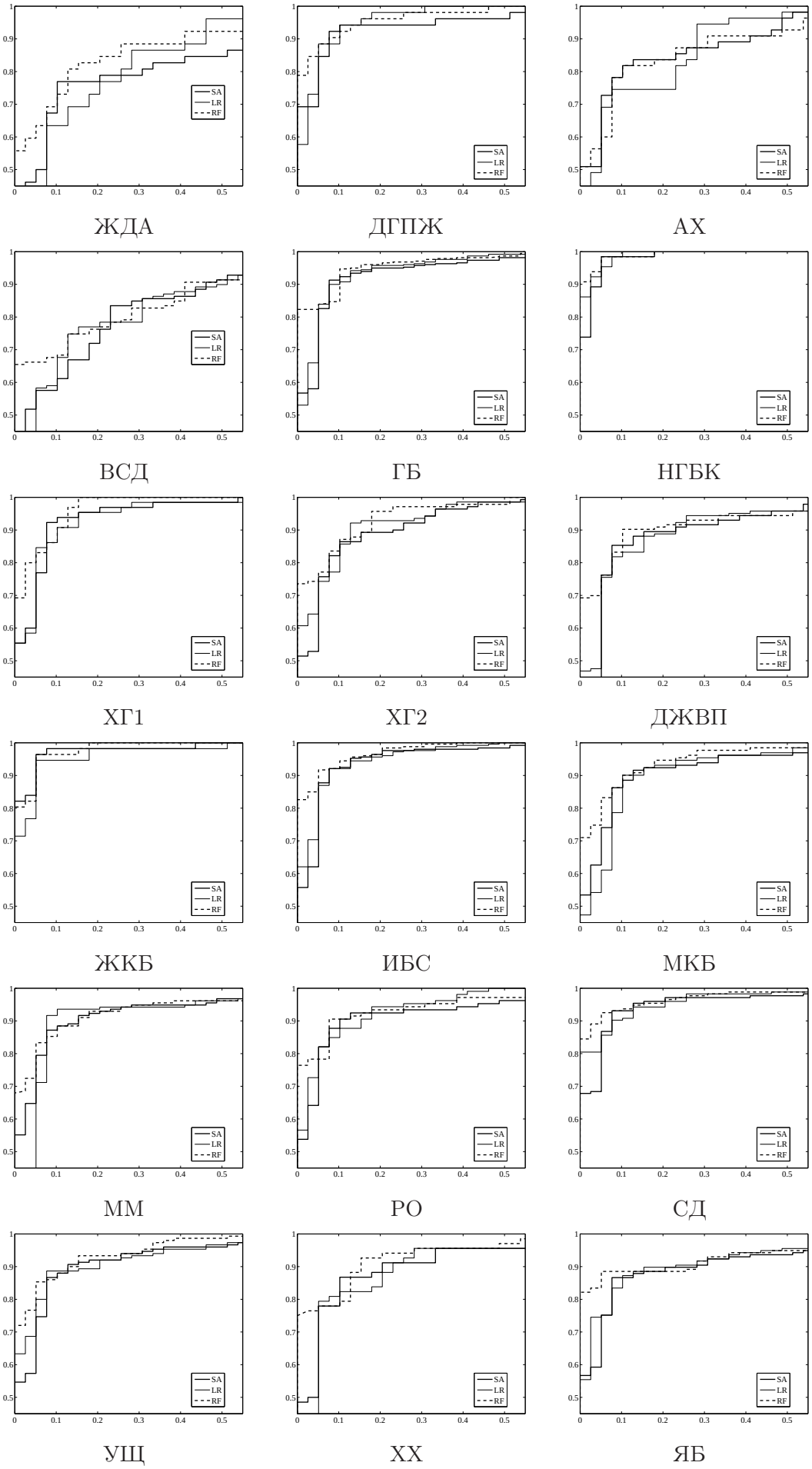


Рис. 19: ROC-кривые трёх методов классификации по всем заболеваниям.

	SA			LR			RF		
	С(Ч=95)	С=Ч	AUC	С(Ч=95)	С=Ч	AUC	С(Ч=95)	С=Ч	AUC
НГБК	94,87	85,69	98,86	94,87	80,93	99,29	94,87	56,57	**99,41**
ЖКБ	94,87	93,73	98,40	82,05	93,98	97,39	94,87	87,34	**98,63**
ИБС	87,18	86,50	96,12	82,05	86,55	96,65	87,18	64,50	**98,14**
СД	87,18	82,08	96,29	79,49	81,13	96,89	84,62	53,86	**97,68**
ХГ1	84,62	93,40	95,78	84,62	92,84	95,46	87,18	87,24	**97,59**
ДГПЖ	66,67	97,27	95,66	84,62	96,99	96,99	84,62	95,00	**97,39**
ГБ	74,36	93,93	95,16	82,05	94,04	95,71	84,62	88,72	**96,95**
ХГ2	66,67	91,66	93,50	66,67	89,51	94,51	82,05	84,78	**95,81**
УЩ	66,67	91,82	93,62	64,10	89,70	94,15	69,23	69,35	**95,80**
МКБ	66,67	97,14	93,89	71,79	97,02	93,23	76,92	95,26	**95,54**
РО	56,41	94,53	93,57	74,36	93,52	95,38	69,23	89,05	**95,38**
ХХ	66,67	92,72	92,46	71,79	90,88	91,40	71,79	83,42	**94,78**
ММ	53,85	91,19	93,67	56,41	90,75	92,44	66,67	77,24	**94,55**
ДЖВП	53,85	92,77	92,11	64,10	91,86	92,04	48,72	79,69	**93,66**
ЯБ	43,59	95,05	91,88	51,28	94,35	92,80	35,90	88,52	**93,22**
АХ	51,28	91,70	**90,26**	64,10	90,61	89,74	46,15	78,41	90,16
ЖДА	23,08	91,24	83,97	53,85	89,07	85,21	30,77	79,75	**89,00**
ВСД	35,90	90,54	85,32	30,77	89,55	84,84	30,77	61,88	**87,21**

Таблица 3: Показатели качества диагностики на контрольной выборке

лов (2T), приращения интервалов и амплитуд (4RT) и приращения интервалов, амплитуд и их отношений (6RTA). При использовании только двух символов, кодирующих приращения интервалов (2T), уже достигается неплохой результат (AUC в среднем 87%), но при использовании только приращений амплитуд (2R) AUC снижается на 5–26%. Совместное использование приращений интервалов и амплитуд при четырёх-символьном кодировании (4RT) даёт небольшой, но значимый прирост качества, в среднем около 2%. Дополнительное использование отношений амплитуд к интервалам при шести-символьном кодировании (6RTA) увеличивает AUC в среднем ещё на 6%.

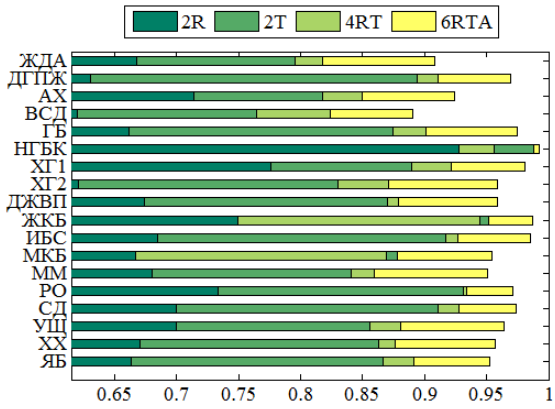


Рис. 20: Значения AUC при использовании 2R, 2T, 4RT и 6RTA кодирований.

6.7 Зависимость AUC от длины кардиосигнала

Для выяснения границ применимости технологии информационного анализа электрокардиосигналов была исследована зависимость величины AUC от количества кардиоциклов N. Для этого при $N = 10, 25, 75, 150, 300, 600$ был построен случайный лес из 100 деревьев по признакам – частотам 2-грамм и измерено значение AUC на тестовой выборке кросс-валидации. Результаты показаны на рис. 21. Увеличение числа кардиоциклов от 300 до 600 незначительно повышает качество классификации (на 1%).

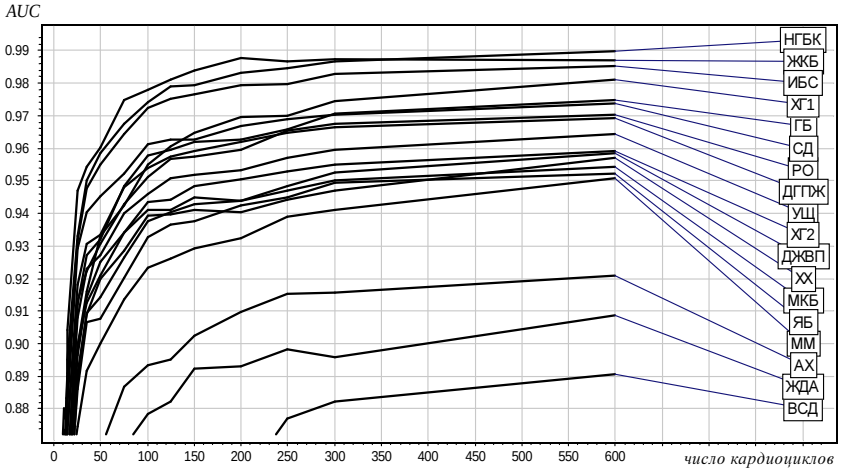


Рис. 21: Зависимости AUC от длины кардиосигнала для различных заболеваний.

7 Заключение

Технология информационного анализа электрокардиосигналов — это совокупность методов регистрации и обработки кардиосигналов с целью получения диагностических эталонов физиологических состояний нормы и заболеваний внутренних органов. В работе проверены статистические гипотезы о неслучайном характере вариаций интервалов и амплитуд кардиоциклов и о взаимосвязи этих вариаций с заболеваниями. Исследован вопрос о возможности диагностики заболеваний по ЭКГ: разработан легко интерпретируемый, практически не подверженный переобучению синдромный алгоритм; достигнуты высокие уровни чувствительности и специфичности для выбранных заболеваний. Обоснован учет совместной вариабельности интервалов, амплитуд и их отношений при кодировании ЭКГ-сигнала, и показано, что для получения высокого качества диагностики достаточно длины кардиосигнала, равной 300. Некоторые из результатов работы вошли в статью [51].

В процессе разработки технологии информационного анализа электрокардиосигналов затронута важная проблема — характер семантики сигналов, генерируемых сердцем. Возможность получения высокоспецифичных кодовых эталонов нормы и заболеваний внутренних органов свидетельствует о том, что кардиосигналы несут во внутреннюю среду организма человека информацию, в которой присутствует семантика здоровья и болезней. Высокое качество классификации для всех выбранных заболеваний позволяет говорить об использовании этой технологии в качестве нового метода диагностики многих заболеваний внутренних органов по одной электрокардиограмме и разработки на её основе принципиально нового класса диагностических систем, не имеющих аналогов в мировой практике.

Список литературы

[1] Баевский Р. М., Иванов Г. Г. Вариабельность сердечного ритма: теоретические аспекты и возможности клинического применения. *Ультразвуковая и функциональная диагностика.* 2001. № 3. С. 108–127.

[2] Баевский Р. М., Иванов Г. Г., Чирейкин Л. В., Гаврилушкин А. П., Довгалевский П. Я., Кукушкин Ю. А., Миронова Т. Ф., Прилуцкий Д. А., Семенов Ю. Н., Федоров В. Ф., Флейшман А. Н., Медведев М. М. Анализ вариабельности сердечного ритма при использовании различных электрокардиографических систем (методические рекомендации). *Вестник аритмологии.* 2001. № 24. С. 65–87.

[3] Успенский В. М. Информационная функция сердца. Теория и практика диагностики заболеваний внутренних органов методом информационного анализа электрокардиосигналов. М.: Экономика и информатика, 2008. 116 с.

[4] Успенский В. М. Информационная функция сердца. *Клиническая медицина.* 2008. Т. 86. № 5. С. 4–13.

[5] Uspenskiy V. M. Information Function of the Heart. Biophysical substantiation of technical requirements for electrocardioblock registration and measurement of electrocardiosignals parameters acceptable for information analysis to diagnose internal diseases. In: *Joint International IMEKO TC1+TC7+TC13 Symposium.* August 31–September 2, 2011, Jena, Germany.

[6] Uspenskiy V. M. Information Function of the Heart. A Measurement Model. In: *Measurement 2011: 8-th International Conference*. Smolenice, Slovakia, April 27–30, 2011. Pp. 383–386.

[7] Uspenskiy V. M. Diagnostic System Based on the Information Analysis of Electrocardiogram. In: *Proceedings of MECO 2012. Advances and Challenges in Embedded Computing.* Bar, Montenegro, June 19–21, 2012. Pp. 74–76.

[8] Успенский В. М., Кравченко Ю. Г., Павловский К. П., Авербах Ю. И. Устройство экспресс-диагностики заболеваний внутренних органов и онкопатологии. Патент на изобретение № 2159574 от 27 ноября 2000 г.

[9] Успенский В. М. Способ диагностики болезней неинфекционной этиологии. Патент на изобретение № 2157093 от 10 октября 2000 г.

[10] Успенский В. М. Способ диагностики заболеваний внутренних органов неинфекционной природы на любой стадии их развития. Патент на изобретение № 2163088 от 20 февраля 2001 г.

[11] Успенский В. М. Способ суточного кардиомониторирования для определения наличия и активности заболеваний человека неинфекционной природы. Патент на изобретение № 2211658 от 10 сентября 2003 г.

[12] Успенский В. М. Способ диагностики заболеваний внутренних органов. Патент на изобретение № 2407431 от 27 декабря 2010 г.

[13] Mika P. Tarvainen, Juha-Pekka Niskanen Kubios HRV version 2.1. User's guide. July 6, 2012.

[14] Andreas Voss, Steffen Schulz, Rico Schroeder, Mathias Baumert, and Pere Caminal Methods derived from nonlinear dynamics for analysing heart rate variability. *Phil. Trans. R. Soc. A.*. 2009. Vol. 367.Issue 1887. Pp. 277–296.

[15] J. Kurths, A. Voss, P. Saparin, A. Witt, H.J. Kleiner, and N.Wessel Quantitative analysis of heart rate variability *Chaos* 1995. Vol. 5. No 1. Pp. 88–94.

[16] Albert C.-C. Yang, Shu-Shya Hseu, Huey-Wen Yien, Ary L. Goldberger and C.-K. Peng Linguistic Analysis of the Human Heartbeat Using Frequency and Rank Order Statistics *Physical review letters* 2003. Vol. 90. No 10.

[17] Camillo Cammarota, Enrico Rogora Time reversal, symbolic series and irreversibility of human heartbeat *Chaos, Solitons & Fractals* 2007. Vol. 32. No 5. Pp. 1649–1654.

[18] U. Parlitz, S. Berg, S. Luther, A. Schirdewan, J. Kurths and N. Wessel Classifying cardiac biosignals using ordinal pattern statistics and symbolic dynamics *Computers in Biology and Medicine* 2012. Vol. 42. No 3. Pp. 319–327.

[19] World Health Organization. http://www.who.int/.

[20] Salem A.-B.M., Revett K., El-Dahshan E.-S.A. Machine learning in electrocardiogram diagnosis. *Computer Science and Information Technology, 2009. IMCSIT '09. International Multiconference on.* 12-14 October 2009. Pp. 429–433.

[21] Edited by Wilbert S. Aronow Cardiac Arrhythmias - Mechanisms, Pathophysiology, and Treatment. Publisher: InTech, 2014. 160 pages.

[22] Saleha Samad, Shoab A. Khan, Anam Haq, and Amna Riaz Classification of Arrhythmia. *International Journal of Electrical Energy*. March 2014. Vol. 2. No. 1.

[23] Mi Hye Song, Jeon Lee, Sung Pil Cho, Kyoung Joung Lee, and Sun Kook Yoo Support Vector Machine Based Arrhythmia Classification Using Reduced Features. *International Journal of Control, Automation, and Systems*. December 2005. Vol. 3. No. 4. Pp. 571–579.

[24] Narendra Kohli and Nishchal K. Verma Arrhythmia classification using SVM with selected features. *International Journal of Engineering, Science and Technology*. 2011. Vol. 3. No. 8. Pp. 122–131.

[25] Emina Alickovic, Abdulhamit Subasi Medical Decision Support System for Diagnosis of Cardiovascular Diseases using DWT and k-NN.

[26] Abhinav Vishwa, Mohit K. Lal, Sharad Dixit, Dr. Pritish Vardwaj Clasification Of Arrhythmic ECG Data Using Machine Learning Techniques. *International Journal of Interactive Multimedia and Artificial Intelligence*. 2011. Vol. 1. No. 4. Pp. 67–70.

[27] Palash Mondal, Kalyani Mali Cardiac Arrhythmias Classification using Decision Tree. *International Journal of Advanced Research in Computer Science and Software Engineering*. January, 2015. Vol. 5. Issue 1. Pp. 540–542.

[28] MIT-BIH Arrhythmia Database. Available at URL: http://physionet.ph.biu.ac.il/physiobank/database/mitdb/

[29] Li Sun, Yanping Lu., Member, IEEE, Kaitao Yang, and Shaozi Li ECG Analysis Using Multiple Instance Learning for Myocardial Infarction Detection. *IEEE Transactions on Biomedical Engineering*. December, 2012. Vol. 59. No. 12. Pp. 3348–3356.

[30] R. S. Wright, J. L. Anderson, C. D. Adams, C. R. Bridges, D. E. Casey, Jr., S. M. Ettinger, F. M. Fesmire, T. G. Ganiats, H. Jneid, A. M. Lincoff, E. D. Peterson, G. J. Philippides, P. Theroux, N. K. Wenger, J. P. Zidar, E. M. Antman, R. M. Califf, W. E. Chavey, J. S. Hochman, and T. N. Levin 2011 ACCF/AHA focused update of the Guidelines for the Management of Patients with Unstable Angina/Non–ST-Elevation Myocardial Infarction (updating the 2007 guideline): A report of the American College of Cardiology Foundation/American Heart Association Task Force

on Practice Guidelines developed in collaboration with the American College of Emergency Physicians, Society for Cardiovascular Angiography and Interventions, and Society of Thoracic Surgeons. *J. Amer. Coll. Cardiol.*. 2011. Vol. 57. No. 19. Pp. 1920–1959.

[31] Getie Zewdie, Momiao Xiong Fully Automated Myocardial Infarction Classification using Ordinary Differential Equations. aiXiv: 1410.6984. October, 2014.

[32] The PTB Diagnostic ECG Database. Available at URL: http://www.physionet.org/physiobank/database/ptbdb/

[33] Б. Л. Мультановский, Л. А. Лещинский, Ю. Л. Кузелин Влияние артериальной гипертензии на частотные показатели вариабельности сердечного ритма по данным суточного мониторирования электрокардиограммы. *Вестник аритмологии.* 2005. № 40. С. 39–44.

[34] Melillo P., Izzo R., Orrico A., Scala P., Attanasio M., Mirra M., et al. Automatic Prediction of Cardiovascular and Cerebrovascular Events Using Heart Rate Variability Analysis. *PLoS ONE*. March 20, 2015. Vol. 10. No. 3.

[35] К. М. Николин Синдром обструктивного сонного апноэ, возможности функциональной диагностики. *Вестник аритмологии.* 2004. № 36. С. 10–17.

[36] Philip de Chazal, Conor Heneghan, Elaine Sheridan, Richard Reilly, Philip Nolan, and Mark O'Malley Automated Processing of the Single-Lead Electrocardiogram for the Detection of Obstructive Sleep Apnoea. *IEEE Transactions on Biomedical Engineering*. June, 2003. Vol. 50. No. 6. Pp. 686–696.

[37] A. Travaglini, C. Lamberti, J. DeBie, M. Ferri Respiratory Signal Derived from Eight-lead ECG. *Computers in Cardiology*. Piscataway, NJ: IEEE Press. 1998. Vol. 25. Pp. 65—68.

[38] F. Roche, V. Pichot, E. Sforza, I. Court-Fortune, D. Duverney, F. Costes, M. Gare, J-C. Barthélémy Predicting sleep apnoea syndrome from heart period: a time-frequency wavelet analysis. *European Respiratory Journal*. 2003. Vol. 22. No. 6. Pp. 937—942.

[39] A. H. Khandoker, C. K. Karmakar, M. Palaniswami Screening Obstructive Sleep Apnoea Syndrome from Electrocardiogram Recordings Using Support Vector Machines. *Computers in Cardiology*. 2007. Vol. 34. Pp. 485—488.

[40] Мучник И. Б., Мучник Р. Б. Алгоритмы формирования языка для описания экспериментальных кривых. *Автоматика и телемеханика*. 1973. выпуск 5. С. 86–98.

[41] Черкай А. Д., Власов Ю. А. Лингвистический анализ ритма сердца — проблемы временной организации живых систем. Отделение физиологии Академии наук СССР. М.: Наука, 1979. С. 62–70.

[42] Маннинг К. Д., Рагхаван П., Шютце Х. Введение в информационный поиск. М.: Вильямс, 2011.

[43] Torshin I. Yu. The study of the solvability of the genome annotation problem on sets of elementary motifs. *Pattern Recognition and Image Analysis*. 2011. Vol. 21, Issue 4. Pp. 652–662.

[44] Good P. I. Permutation, Parametric, and Bootstrap Tests of Hypotheses. Springer Science & Business Media, 2006. 336 p.

[45] Журавлёв Ю. И., Рязанов В. В., Сенько О. В. «Распознавание». Математические методы. Программная система. Практические применения. М.: ФАЗИС, 2006. 176 с.

[46] Мерков А. Б. Распознавание образов. Введение в методы статистического обучения. Едиториал УРСС, 2011. 256 с.

[47] Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning, 2nd edition. Springer, 2009. 533 p.

[48] Breiman Leo Random Forests. *Machine Learning*. October 2001. Vol. 45. Issue 1. Pp. 5–32.

[49] Breiman Leo Bagging predictors. *Machine Learning*. August 1996. Vol. 24. Issue 2. Pp. 123–140.

[50] Ho Tin The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. August 1998. Vol. 20, Issue 8. Pp. 832–844.

[51] Uspenskiy V., Vorontsov K., Tselykh V., Bunakov V. Information Function of the Heart: Discrete and Fuzzy Encoding of the ECG-Signal for Multidisease Diagnostic

System. In *Advanced Mathematical and Computational Tools in Metrology and Testing X, Series on Advances in Mathematics for Applied Sciences*. 2015. World Scientific. Singapore. Vol. 86. Pp. 377–384.