



Магистерская программа  
«Логические и комбинаторные методы анализа данных»

Магистерская диссертация  
**«Ансамбль алгоритмов для определения ишемической  
болезни сердца по электрокардиограмме»**

Работу выполнила:  
**Гурьянова Валерия Николаевна**

Научный руководитель:  
профессор, д.ф.-м.н.  
**Дьяконов Александр Геннадьевич**

# Содержание

<b>1</b>	<b>Введение</b>	<b>4</b>
<b>2</b>	<b>Постановка задачи и описание данных</b>	<b>5</b>
<b>3</b>	<b>Предобработка данных</b>	<b>5</b>
<b>4</b>	<b>Алгоритмы, входящие в состав ансамбля</b>	<b>6</b>
4.1	Алгоритм, основанный на признаках, полученных из HRV-сигнала . . . . .	7
4.2	Алгоритм, основанный на 3-х различных группах признаков . . . . .	13
4.3	Алгоритм, основанный на анализе сегментов сигнала вокруг его R-пиков	15
4.4	Алгоритм, основанный на вейвлет-преобразовании сигнала . . . . .	16
4.5	Алгоритм, основанный на биспектральном разложении сигнала . . . . .	19
<b>5</b>	<b>Методы построения ансамблей алгоритмов</b>	<b>21</b>
5.1	Голосование по большинству . . . . .	21
5.2	EM-алгоритм . . . . .	21
<b>6</b>	<b>Критерии качества, используемые для сравнения реализованных алгоритмов</b>	<b>22</b>
6.1	Качество по пациентам . . . . .	22
6.2	F-мера . . . . .	23
6.3	ROC-AUC . . . . .	24
<b>7</b>	<b>Результаты экспериментального исследования</b>	<b>24</b>
7.1	Характеристики программной реализации . . . . .	24
7.2	Метод оценки алгоритмов и подбора параметров . . . . .	24
7.3	Полученные результаты . . . . .	25
<b>8</b>	<b>Заключение</b>	<b>29</b>

## **Аннотация**

Ишемическая болезнь сердца является ведущей причиной смерти в мире. Данная болезнь может протекать долгое время бессимптомно и с течением времени может прогрессировать и закончиться смертью. В современных условиях с помощью специального оборудования стало возможно снять кардиограмму в домашних условиях. В данной работе ставилась задача определения ишемической болезни по электрокардиограммам, полученных с помощью данного оборудования. Проводилось исследование различных подходов для классификации электрокардиограмм. Были предложены новые подходы и модификации существующих. Кроме того были предложены способы ансамблирования реализованных алгоритмов. Наилучшее качество показал ансамбль, реализованный с помощью EM-алгоритма.

# 1 Введение

Ишемическая болезнь сердца (ИБС) – это такое состояние, при котором нарушается кровообращение мышцы сердца, когда к некоторым его участкам периодически и при определенных обстоятельствах поступает недостаточное количество крови. Проявлениями ИБС могут быть стенокардия, острая коронарная недостаточность со смертельным исходом, инфаркт миокарда и нарушение сердечного ритма. [1] По данным Всемирной организации здравоохранения [2] ишемическая болезнь сердца является ведущей причиной смертей в мире.

На начальных стадиях заболевания у большинства людей не проявляется никаких симптомов данного заболевания. Однако со временем болезнь чаще всего прогрессирует и может закончиться смертью [1]. Очень важно своевременно выявить заболевание, чтобы замедлить ход развития болезни и не допустить смерти.

Традиционно ишемическая болезнь сердца определяется при помощи специалистов и ряда исследований. Исследования включают в себя тестирование под нагрузкой (если больной не в состоянии его выполнить, то применяется ввод лекарственных препаратов, которые увеличивают нагрузку на сердце), анализы крови, мониторинг артериального давления, электрокардиограмму, эхокардиограмму и прочее [1]. Следует отметить, что эти исследования занимают существенное время пациента, а также требуют высокой квалификации специалиста, который будет их проводить. Так как специалистов крайне мало, а число потенциальных больных растет с каждым годом, то задача автоматического определения ишемической болезни сердца является крайне актуальной в настоящее время. В настоящее время крайне актуальной является задача создания устройства, которое поможет определять заболевание или его вероятность в домашних условиях. Такие устройства позволят направить человека к врачу, в случае высокой вероятности наличия у него ИБС.

Электрокардиограмма (ЭКГ) – это сигнал, отображающий электрическую активность сердца. Он может быть измерен при помощи электродов, прикрепленных к конечностям или грудной клетке [3]. ЭКГ является одним из наиболее доступных средств диагностики сердечных заболеваний в настоящее время из-за ее неинвазивности и низкой стоимости. В настоящее время существует множество различных исследований, которые показывают, что ЭКГ может быть использована для определения ишемической болезни сердца [4] [5]. Компания CardioQvark создала устройство, которое представляет собой насадку на смартфон, позволяющую делать замеры ЭКГ в бытовых условиях. Замеры отправляются в облако и анализируются там, а затем результаты отправляются пациенту и лечащему врачу.

В данной работе исследовались различные методы, которые позволяют выявлять болезни сердца на основе ЭКГ, и их применимость для определения ишемической болезни сердца по ЭКГ, полученном от устройства, созданного компанией CardioQvark. Были предложены модификации этих методов, которые позволяют улучшить качество классификации для данной задачи. Также было проведено исследование применимости нового метода для автоматического определения ишемической болезни сердца. С целью улучшения качества классификации был построен ансамбль из реализованных алгоритмов.

Данная работа состоит из следующих разделов: постановка задачи и описание данных, описание предобработки данных, описание алгоритмов, входящих в состав ансамбля, описание методов построения ансамблей алгоритмов, описание критериев качества, по которым сравниваются алгоритмы, результаты экспериментальных исследований и заключение.

## 2 Постановка задачи и описание данных

Целью данной работы является построение алгоритма, который показывает наилучшее качество при определении ишемической болезни сердца по электрокардиограмме, полученной с помощью устройства от компании CardioQvark. Следует отметить, что эталонного решения для данной задачи не существует.

При решении данной задачи использовались две выборки: выборка, по которой проводилась кросс-валидация и обучение алгоритма, и выборка, которая использовалась для тестирования.

Выборка, которая использовалась для обучения алгоритмов состоит из 1798 кардиограмм. При этом она содержит 1055 кардиограмм здоровых пациентов и 743 кардиограммы больных пациента. Частота дискретизации 1000 Гц.

Выборка, которая использовалась для тестирования алгоритмов состоит из 1567 кардиограмм. При этом она содержит 973 кардиограммы здоровых пациентов и 594 кардиограммы больных пациентов. Частота дискретизации 1000 Гц.

## 3 Предобработка данных

Данные, полученные от электрокардиографа, часто подвержены шуму, поэтому необходима предобработка данных. Для избавления данных от шума используются различные фильтры. Фильтр — это обрабатывающая система, алгоритм или метод, реализованный программно или аппаратно, который используется, чтобы изменить заданный

сигнал специальным образом [3].

Фильтр нижних (верхних) частот – это фильтр, который используется, чтобы удалить высокочастотный (низкочастотный) шум с минимальными потерями компонент сигнала в указанной полосе пропускания [3].

Каждый фильтр характеризуется величиной, которая называется амплитудно-частотная характеристика. Амплитудно-частотная характеристика показывает зависимость амплитуды выходного сигнала от частоты входного сигнала [3].

Фильтр Баттерворта (нижних и верхних частот) является одним из наиболее часто используемых фильтров по причине простоты его реализации, а также из-за максимально гладкой амплитудно-частотной характеристики, которая позволяет не слишком сильно изменять вид исходного сигнала [3].

Медианный фильтр часто используется для извлечения тренда, который был получен из-за помех в приборе. Медианный фильтр определяется следующим образом.

Пусть  $w = 2h + 1$  – заданный размер окна, где  $h$  – целое число,  $x$  – исходный сигнал,  $i$  – индекс элемента сигнала. `median` - статистика, такая что половина элементов выборки меньше ее, и половина элементов выборки больше ее. Тогда выходной сигнал  $y$  определяется как:

$$y_i = \text{median}(x_{i-h}, x_{i-h+1}, \dots, x_{i+h}).$$

Проведенная предобработка данных включает в себя следующие шаги:

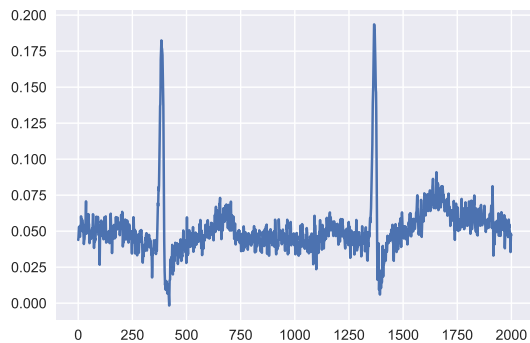
- Применение фильтра Баттерворта второго порядка верхних частот с частотой подавления сигнала 0.3 Гц.
- Применение фильтра Баттерворта второго порядка верхних частот с частотой подавления сигнала 0.3 Гц.
- Извлечение тренда при помощи медианного фильтра.
- Вычитание тренда из полученного сигнала.

В данной работе использовались реализации указанных выше фильтров из модуля `signal` библиотеки `scipy` [6].

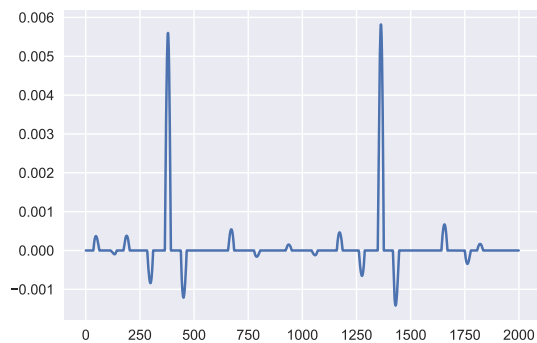
На Рис. 1 изображен пример ЭКГ-сигнала до и после предобработки.

## 4 Алгоритмы, входящие в состав ансамбля

Ниже приведены описания алгоритмов, которые были использованы для построения ансамбля.



Сигнал до предобработки



Сигнал после предобработки

Рис. 1: Вид сигнала до и после предобработки.

#### 4.1 Алгоритм, основанный на признаках, полученных из HRV-сигнала

Данный алгоритм был описан в статье [4] и был использован для обнаружения ишемической болезни сердца.

В ЭКГ-сигнале можно выделить R-пики, которые соответствуют пульсу человека [3]. Пример того, как выглядят R-пики в ЭКГ-сигнале можно увидеть на Рис. 2.

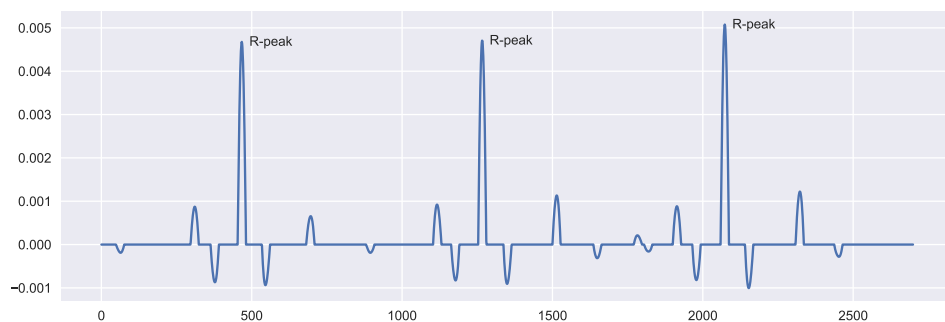


Рис. 2: R-пики в ЭКГ сигнале.

Алгоритм обнаружения R-пигов подробно описан в [7] и его общий принцип состоит в следующем.

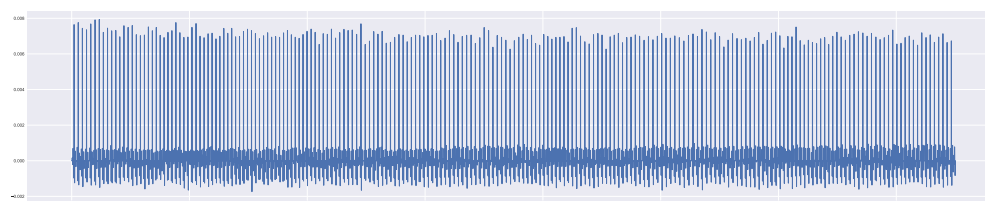
- Выделяем только те пики, в окрестности 200 мс которых нет пиков больше данного.
- Если такие пики нашлись, то проверяем, содержит ли ЭКГ-сигнал положительный и отрицательный наклон.
- Если пик больше чем, некоторый заданный порог, то считаем его R-пиком.

Для вычисления R-пику использовалась реализация указанного выше алгоритма из библиотеки BioSPPy [8].

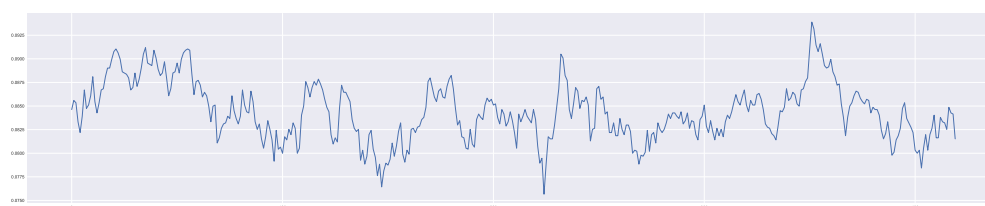
По ЭКГ-сигналу может быть построен сигнал, который называется вариабельность сердечного ритма (HRV-сигнал). Он вычисляется следующим образом.

- Вычисляются R-пики.
- Измеряются интервалы между двумя R-пиками (RR-интервалы).
- Каждая величина RR-интервала преобразуется в  $60/RR$ .

На Рис. 3 изображен ЭКГ-сигнал и соответствующий ему HRV-сигнал.



ЭКГ-сигнал



HRV-сигнал

Рис. 3: ЭКГ и HRV-сигнал.

На основе HRV-сигнала строятся различные группы признаков.

В первую группу признаков входят различные энтропийные признаки, которые показывают меру непредсказуемости в сигнале. Используются следующие виды энтропий: приближенная энтропия, выборочная энтропия и энтропия Шеннона. Про каждый вид энтропии подробно описано ниже.

Приближенная энтропия вычисляется следующим образом. Здесь  $x$  — HRV-сигнал длины  $N$ .

- Необходимо зафиксировать целое  $m$  и действительное  $r$ .
- На основе сигнала составим набор векторов следующего вида:  $\overline{x}_i^m = (x_i, x_{i+1}, \dots, x_{i+m-1})$ , где  $i \in [0, N - m]$ .



- Определим величины  $C_i^m(r)$  следующим образом:

$$C_i^m(r) = \frac{|\overline{x_k^m} : d(\overline{x_i^m}, \overline{x_k^m}) \leq r, k \in [0, N - m]|}{N - m + 1},$$

$$d(\overline{x_i^m}, \overline{x_k^m}) = \max_{a \in [0, m-1]} \text{abs}(\overline{x_i^m}(a) - \overline{x_k^m}(a)), \quad (1)$$

где  $\overline{x_i^m}(a)$  – компонента  $a$  вектора  $\overline{x_i^m}$ ,  $\text{abs}(v)$  – абсолютное значение величины  $v$ .

- Определим величины  $\Phi^m(r)$  как:

$$\Phi^m(r) = (N - M + 1)^{-1} \sum_{i=1}^{N-m+1} \log(C_i^m(r)).$$

- Приближенная энтропия (ApEn) определяется как

$$\text{ApEn}(r) = \Phi^m(r) - \Phi^{m+1}(r).$$

В данной работе приближенная энтропия реализовывались для  $m = 10$  и  $r = 0.2\text{std}(x)$ ,  $\text{std}(x)$  – стандартное отклонение и определяется по формуле.

$$\text{std}(x) = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}, \quad (2)$$

где  $x_i$  –  $i$ -ая компонента сигнала,  $N$ -число элементов в сигнале,  $\mu$  – среднее сигнала и определяется по формуле:

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i. \quad (3)$$

Выборочная энтропия вычисляется следующим образом.

- Формируются векторы  $\overline{x_i^m}$  длины  $m$  и векторы  $\overline{x_i^{m+1}}$  длины  $m + 1$  аналогичные тем, которые были сформированы в приближенной энтропии.
- Вычисляются величины  $A$  и  $B$  следующим образом:

$$A(r) = |\overline{(x_k^{m+1}, x_l^{m+1})} : d(\overline{x_k^{m+1}}, \overline{x_l^{m+1}}) \leq r, 0 \leq k \leq l \leq N - m - 1|,$$

$$B(r) = |\overline{(x_k^m, x_l^m)} : d(\overline{x_k^m}, \overline{x_l^m}) \leq r, 0 \leq k \leq l \leq N - m|,$$

где  $d$  определяется согласно формуле (1).

- Выборочная энтропия (SampEn) определяется как

$$\text{SampEn}(r) = -\log \frac{A(r)}{B(r)}.$$

В данной работе выборочная энтропия определялась при  $m = 10$ ,  $r = 0.2\text{std}(x)$ .

Энтропия Шеннона (ShanEn), вычисляется как

$$\text{ShanEn} = \sum_{f=1}^k p_f \log p_f,$$

где  $k$  – число различных элементов в сигнале  $x$ ,  $p_f$  – частота элемента  $f$  в сигнале  $x$ .

Следующая группа признаков основана на графике повторений. Данный график показывает частоту и продолжительность повторений в сигнале. Элемент  $(i, j)$  данного графика определяется следующим образом

$$R(i, j) = \begin{cases} 1, & \text{если } ||x_i - x_j|| < \varepsilon \\ 0, & \text{иначе} \end{cases},$$

где  $x$  – HRV-сигнал.

На основе данного графика вычисляются следующие признаки,  $N$  – число элементов в HRV-сигнале,  $l_{d \min}$  – минимальная длина диагонали на графике,  $l_{v \min}$  – минимальная длина вертикальной линии,  $l_{d \max}$  – максимальная длина диагонали на графике,  $l_{v \max}$  – максимальная длина вертикальной линии:

- Плотность точек на графике (REC)

$$\text{REC} = \frac{1}{N^2} \sum_{i,j=0}^N R(i, j).$$

- Процент точек, которые формируют диагональные линии (DET)

$$\text{DET} = \frac{\sum_{l=l_{d \min}}^{l_{d \max}} lP(l)}{\sum_{i,j}^N R(i, j)},$$

$P(l)$  – число диагоналей длины  $l$ .

- Средняя длина диагоналей ( $L_{\text{mean}}$ )

$$L_{\text{mean}} = \frac{\sum_{l=l_{d \min}}^{l_{d \max}} lP(l)}{\sum_{l=l_{d \min}}^{l_{d \max}} P(l)}.$$

- Энтропия диагональных линий ( $EN_d$ )

$$EN_d = - \sum_{l=l_{d \min}}^{l_{d \max}} p_l \log p_l,$$

где  $p_l$  – частота диагональных линий длины  $l$ .

- Энтропия вертикальных линий ( $EN_v$ )

$$EN_v = - \sum_{l=l_{v \min}}^{l_{v \max}} p_{vl} \log p_{vl},$$

где  $p_{vl}$  – частота вертикальных линий длины  $l$ .

Для вычисления указанных признаков в данной работе использовалась библиотека PyRQA [9].

Еще одна группа признаков, которая используется при данном подходе, – это группа признаков, основанных на графике Пуанкаре. График Пуанкаре строится следующим образом: для сигнала  $x = (x_1, x_2, \dots, x_N)$  на графике отображаются точки

$$(x_1, x_2), (x_2, x_3), \dots, (x_i, x_{i+1})$$

и так далее. В данном случае в качестве сигнала берется сигнал, который состоит из RR-интервалов.

Используются следующие признаки:

- Стандартное отклонение (формула (2)) расстояний от точек графика до прямой  $y = x$ . Данный признак описывает локальную вариативность RR-интервалов.
- Стандартное отклонение (формула (2)) расстояний от точек графика до прямой  $y = -x - 2RR_{mean}$ .  $RR_{mean}$  – среднее значение RR-интервалов. Данный признак описывает долговременную вариативность RR-интервалов.

Признаки, основанные на детрентном флуктуационном анализе. Данный метод позволяет определить самозависимость сигнала. Определим следующую кумулятивную сумму:

$$x_{cumsum}(t) = \sum_{i=1}^t (x(i) - \mu),$$

$x$  – сигнал, состоящий из RR-интервалов,  $\mu$  – среднее значение сигнала  $x$  (определяется по формуле (3)).

Данные сегментируются с окном размера  $\Delta n$ . На каждом сегменте для данных находится полином, который наиболее точно их представляет (обычно линейный). Объединение всех таких полиномов образует функцию  $x_{\Delta n}(t)$  являющуюся приближением исходной функции  $x_{cumsum}(t)$ . Далее находится следующая функция:

$$F(\Delta n) = \sqrt{\frac{1}{N} \sum_{t=1}^N [x_{cumsum}(t) - x_{\Delta n}(t)]^2},$$

где  $N$  – длина сигнала, состоящего из RR-интервалов.

Признаком будет являться угол наклона линии  $\log F(\Delta n)$  к  $\log(\Delta n)$ . Подробнее про данный подход написано в [10]. В данной работе для вычисления данного признака использовалась реализация из библиотеки `nolds` [11].

Следующим признаком, который используется при данном подходе является корреляция размерностей. Данный признак является количественной характеристикой траектории сигнала и определяется следующим образом.

- Формируются векторы  $\overline{x_i^m}$  длины  $m$  аналогичные тем, которые были сформированы в приближенной энтропии.
- Величина  $g$  определяется как:

$$g(r) = |\{(\overline{x_k^m}, \overline{x_l^m}) : d(\overline{x_l^m}, \overline{x_k^m}) \leq r, 0 \leq k \leq l \leq N - m - 1\}|.$$

$g$  – это число пар векторов, расстояние между которым меньше либо равно  $r$ .

- Величина  $C(r)$  определяется как

$$C(r) = \frac{g(r)}{N^2},$$

где  $N$  – это длина сигнала  $x$ .

- Корреляция размерностей ( $D2$ ) определяется как

$$D2 = \lim_{r \rightarrow 0} \frac{\log C(r)}{\log(r)}.$$

В данной работе использовалась реализация данного признака из библиотеки `nolds` [11].

В качестве классификатора при данном подходе использовалась логистическая регрессия из пакета `scikit-learn` [12] и градиентный бустинг из библиотеки `xgboost` [13].

## 4.2 Алгоритм, основанный на 3-х различных группах признаков

Данный алгоритм представляет собой смесь 3-х различных признаков пространств, которые ранее были использованы в задачах классификации биомедицинских сигналов.

Первая группа признаков состоит из параметров Йорта [14]. Данные параметры изначально использовались как характеристики электроэнцефалограмм. Электроэнцефалограмма (ЭЭГ) – это сигнал, который показывает биоэлектрическую активность головного мозга [3]. Позже данные параметры стали применяться во многих работах, в том числе и для классификации ЭКГ-сигналов [15]. Данные параметры определяются следующим образом.  $x(i)$  – заданный сигнал,  $\frac{dx(i)}{di}$  – аналог производной для дискретного сигнала, которая определяется следующим образом,

$$\frac{dx(i)}{di} = x(i+1) - x(i), \quad i \in [0, N-2],$$

$\text{var}(x)$  – дисперсия сигнала  $x$  и определяется следующим образом:

$$\text{var}(x) = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2, \quad (4)$$

где  $\mu$  – среднее и определяется по формуле (3).

- Активность (Activity) отображает дисперсию сигнала:

$$\text{Activity} = \text{var}(x(i)).$$

- Мобильность (Mobility) представляет собой среднюю частоту:

$$\text{Mobility} = \sqrt{\frac{\text{var}\left(\frac{dx(i)}{di}\right)}{\text{var}(x(i))}}.$$

- Сложность (Complexity) показывает на изменения в частоте сигнала:

$$\text{Complexity} = \frac{\text{Mobility}\left(\frac{dx(i)}{di}\right)}{\text{Mobility}(x(i))}.$$

Следующая группа признаков представляет собой статистические характеристики сигнала и состоит из следующих признаков.

- Среднее сигнала ( $\mu$  – формула (3)).
- Стандартное отклонение сигнала (формула (2)).
- Минимум и максимум значений сигнала.

- Пусть задана выборка  $x = (x_1, x_2, \dots, x_N)$ , ей соответствует вариационный ряд  $x^{(1)} \leq x^{(2)} \leq \dots x^{(N)}$ . Квантилью порядка  $\alpha$  называется число равное элементу, которому соответствует номер  $N\alpha + 1$  в вариационном ряду. В качестве признаков берутся выборочные квантили сигнала порядка 0.1, 0.25, 0.5, 0.75, 0.9.
- Суммы и суммы квадратов значений сигнала, которые находятся выше/ниже определенных значений квантилей 0.1, 0.25, 0.5, 0.75, 0.9.
- Коэффициент асимметрии:

$$\frac{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^3}{\left(\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2\right)^{3/2}},$$

где  $N$  — длина сигнала.

- Коэффициент эксцесса:

$$\frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^4}{\left(\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2\right)^2} - 3,$$

где  $N$  — длина сигнала.

Следующая группа признаков была предложена В. М. Успенским для определения заболевания по ЭКГ пациента [16]. Для вычисления данных признаков необходимо вычислить амплитуды R-пиков ( $A(n)$ ), расстояния между R-пиками ( $T(n)$ ), а также арктангенс их отношения

$$\alpha(n) = \text{arctg} \frac{A(n)}{T(n)}.$$

Изображение данных величин представлено на Рис. 4.

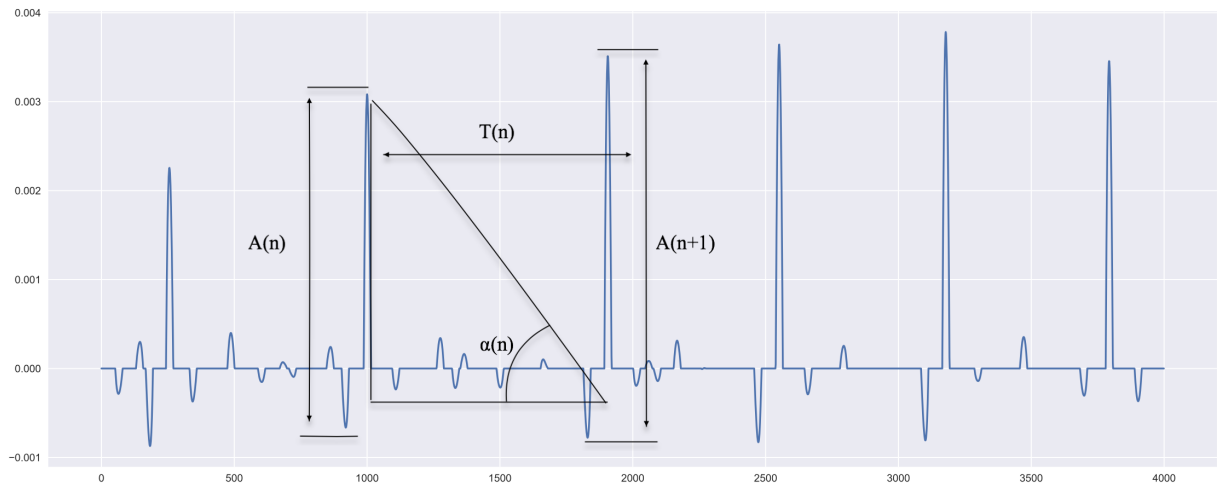


Рис. 4: Величины, используемые при формировании признаков Успенского.

Предполагается, что имеют значения не сами величины  $A(n)$ ,  $T(n)$ , а лишь знаки их приращений. Способ кодирования сигнала, основанный на всех возможных знаках приращений указанных величин представлен в Таблице 1.

	A	B	C	D	E	F
$A(n + 1) - A(n)$	+	-	+	-	+	-
$T(n + 1) - T(n)$	+	-	-	+	+	-
$\alpha(n + 1) - \alpha(n)$	+	+	+	-	-	-

Таблица 1: Кодовое преобразование сигнала

После того как получено кодовое представление сигнала, осуществляется выделение три-грамм. Признаковым пространством является количество вхождений каждой из возможных три-грамм в заданную кодовую последовательность, полученную из сигнала.

В качестве классификатора при данном подходе использовалась логистическая регрессия из пакета `scikit-learn` [12] и градиентный бустинг из библиотеки `xgboost` [13].

### 4.3 Алгоритм, основанный на анализе сегментов сигнала вокруг его R-пику

Данный алгоритм был предложен авторами статьи [17] для определения состояния сердца пациента, при котором возникает необходимость в госпитализации.

Построение признакового пространства для данного подхода осуществляется следующим образом.

- Выделяются окрестности R-пику сигнала: 200 мс до R-пику и 500 мс после.
- В качестве признакового пространства используется усредненная окрестность.

В качестве классификатора использовалась искусственная нейронная сеть. Искусственная нейронная сеть — это модель, построенная по аналогии с принципом функционирования биологических нейронных сетей. Основным принципом построения искусственного нейрона состоит в следующем: каждый нейрон получает входные сигналы по каналам связи, соответствующим каналам от предыдущего нейрона. Для каждого канала связи установлен его вес, а также функция активации. Выход нейрона определяется как:

$$y(x, w) = f\left(\sum_{i=1}^N w_i x_i\right),$$

где  $w_i$  – заданные веса для  $i$ -го канала связи,  $x_i$  – входные данные для  $i$ -го канала связи,  $f$  – функция активации. Совокупность таких нейронов вместе с выделенными входными и выходными нейронами может служить как универсальная аппроксимация произвольной функции [18].

Обучение нейронных сетей происходит с помощью методов стохастической оптимизации, используя технику обратного распространения ошибки.

В нейронных сетях выделяют группу сетей прямого распространения, нейроны и связи между которыми образуют направленный граф без циклов. Большинство таких сетей можно представить как последовательность наборов нейронов, в котором связи могут идти только из нейронов предыдущего набора в нейроны последующего набора. Такие наборы называются слоями.

Одним из типов слоев, которые часто используются в нейронных сетях, является полносвязный слой. В таком слое канал связи идет из каждого нейрона предыдущего слоя в каждый нейрон текущего слоя.

В данной работе рассматривалась следующая архитектура нейронной сети, которая описана в Таблице 2, где  $f_s$  – сигмоидная функция активации. Данная функция активации задается как:

$$f_s(x) = \frac{1}{1 + e^{-x}}.$$

Слой	Входной слой	Полносвязный слой	Полносвязный слой
Параметры	число нейронов=700	число нейронов=90 функция активации: $f_s$	число нейронов=1 функция активации: $f_s$

Таблица 2: Структура нейронной сети

В данной работе нейронная сеть была реализована с помощью библиотек Theano [19] и Lasagne [20], в качестве метода оптимизации использовался алгоритм Adam [18].

#### 4.4 Алгоритм, основанный на вейвлет-преобразовании сигнала

Данный алгоритм, использовался авторами статьи [21] для обнаружения эпилепсии по ЭЭГ-сигналу, а также для идентификации человека по ЭКГ-сигналу.

Вейвлет-преобразование сигнала – это свертка функций  $\Psi(t)$ , называемых вейвлетами, с сигналом. Данные вейвлет-функции должны иметь следующие свойства:

$$\int_{-\infty}^{+\infty} \Psi(t) dt = 0 \quad \int_{-\infty}^{+\infty} |\Psi(t)|^2 dt < \infty.$$



Вейвлет-преобразование позволяет получить сжатие сигнала, которое обладает хорошим качеством восстановленного сигнала [3].

Все вейвлет-функции, используемые в вейвлет-преобразовании могут быть представлены на основе функции-прототипа  $\psi(t)$ , путем масштабирования и переноса. Так в случае дискретного вейвлет-преобразования все вейвлет-функции могут быть записаны как:

$$\psi_{m,n}(t) = \frac{1}{\sqrt{2^m}} \psi(2^{-m}t - n).$$

Параметр сдвига  $n$  показывает, где во временной оси размещен соответствующий вейвлет, а параметр масштабирования  $m$  связан с частотой для заданного вейвлета.

Тогда вейвлет-преобразование может быть записано как:

$$X(m, n) = \int_{-\infty}^{+\infty} x(t) \psi_{m,n}^*(t) dt,$$

где  $\psi^*$  — это функция комплексно сопряженная к  $\psi$ .

В дискретном вейвлет-преобразовании функции  $\psi_{m,n}^*$  могут быть разбиты на 2 составляющие. Это функции, которые отвечают коэффициентам аппроксимации и коэффициентам детализации. В данной работе для представления сигнала используются коэффициенты аппроксимации для нового представления сигнала, а в качестве вейвлет-функций используются Daubechies вейвлеты [22].

Программная реализация используется из библиотеки PyWavelets [23].

После того как были получены коэффициенты вейвлет-преобразования, из сигнала выделяются локальные сегменты. Для этого вдоль сигнала необходимо пройти «окном» некоторой длины  $w$ , с шагом  $s$ . Все элементы, которые попадают в пределы одного окна, записываются в отдельный сегмент. Таким образом, каждый сигнал представляет собой набор локальных сегментов.

Все локальные сегменты, которые входят в обучающую выборку кластеризуются на  $k$  кластеров методом  $k$ -средних. После этого каждый сегмент заменяется номером кластера, которому он принадлежит. Таким образом каждый сигнал представляется в виде текста, состоящего из кодовых слов, соответствующих выделенным кластерам.

При реализации заданного подхода использовались следующие значения параметров:  $w = 100$ ,  $s = 30$ ,  $k = 200$ . Реализация метода  $k$ -средних использовалась из библиотеки scikit-learn.

В обучающей выборке каждый локальный сегмент заменяется на тот кластер, ближе всего к которому он расположен. То есть для локального сегмента  $s_i$  в тестовой выборке кластер  $c$ , которому он принадлежит определяется формулой:

$$c = \underset{j}{\operatorname{argmin}} d(b_j, s_i), \quad d(b_j, s_i) = \sqrt{\sum_{k=1}^w (s_i^k - b_j^k)^2},$$

где  $b_j$  – центр кластера  $j$ ,  $s_i^k (b_j^k)$  –  $k$ -ый элемент локального сегмента  $i$  (кластера  $j$ ).

Таким образом, получив вместо сигнала текст, можно использовать подходы, применимые для текстов. Авторы статьи, которые предложили данное кодирование сигнала, использовали в качестве признакового пространства «мешок слов». Признаковым описанием является число вхождений каждого кодового слова в данный сигнал.

Кроме того, с целью выделения зависимостей между локальными сегментами в сигнале использовались признаки на основе word2vec. Данный подход был предложен компанией Google и он позволяет учитывать контекст слов при обработке текстов, в то же время сокращая размер данных [24]. Word2Vec представляет собой на самом деле два разных метода: Continuous Bag of Words (CBOW) и Skip-gram. В CBOW методе, целью алгоритма является предсказать слово, на основе слов, которые его окружают. В методе Skip-gram целью является предсказать слова, которые окружают данное слово. Оба метода используют нейронную сеть в качестве алгоритма для обучения. В начале обучения каждое слово – это случайный вектор. Во время обучения алгоритм изучает оптимальный вектор для каждого слова, используя метод CBOW или Skip-gram.

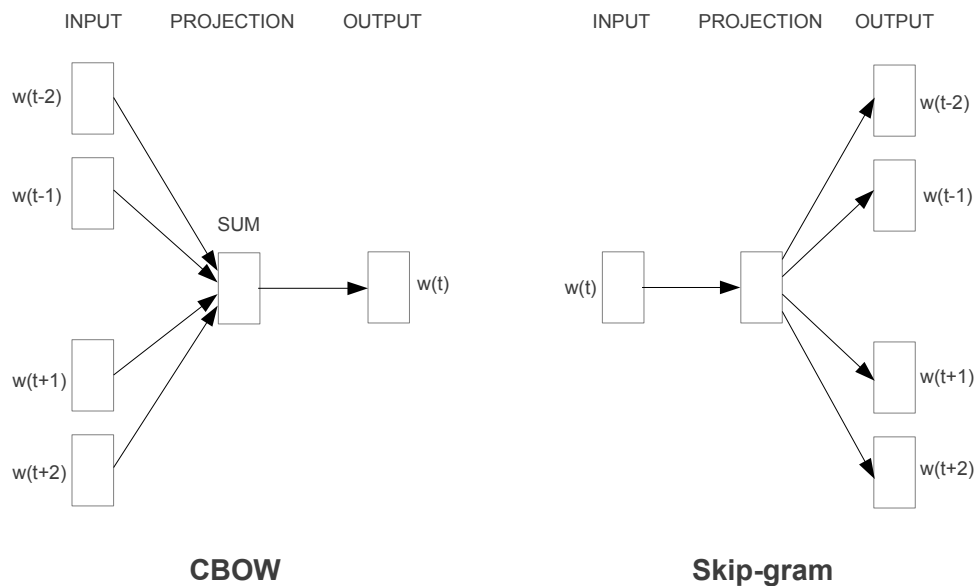


Рис. 5: Основной принцип работы методов Word2Vec [24]

Модель word2vec строилась по обучающей выборке с длиной вектора 80. В качестве признаков использовался вектор, являющийся средним всех векторов, которые входят в текст соответствующий сигналу.

Модель строилась с помощью библиотеки gensim [25].

В качестве алгоритма классификации использовалась логистическая регрессия из библиотеки scikit-learn [12].

## 4.5 Алгоритм, основанный на биспектральном разложении сигнала

Преобразование Фурье [3] – это операция, сопоставляющая одной функции  $x(t)$  другую функцию  $X(f)$  согласно следующей формуле:

$$X(f) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x(t)e^{-itf} dt.$$

Биспектр сигнала – это функция от двух переменных  $f_1$  и  $f_2$ , задающих частоты, выражающаяся следующей формулой [26]:

$$B(f_1, f_2) = X(f_1)X(f_2)X^*(f_1 + f_2),$$

где  $X(f)$  – преобразование Фурье сигнала, а  $X^*(f)$  – комплексное сопряженное к нему. Биспектр сигнала как правило рассчитывается с помощью быстрого преобразования Фурье. Подробное описание алгоритма для нахождения биспектра сигнала можно найти в [27]. При вычислении биспектра сигнала получается двумерная матрица, элементами которой являются комплексные числа.

На основе полученной матрицы, элементы которой можно обозначить как  $a(i, j)$ , можно сопоставить каждому сигналу некоторое изображение. Данное изображение вычисляется следующим образом. Вычисляется новая матрица  $B = ||b_{i,j}||$ , элементы которой равны:

$$b(i, j) = \sqrt{\operatorname{Re}^2 a(i, j) + \operatorname{Im}^2 a(i, j)},$$

где  $\operatorname{Re}$  – обозначает действительную часть комплексного числа, а  $\operatorname{Im}$  обозначает мнимую часть комплексного числа. В качестве изображения используется график линий уровня матрицы  $B$ . Линия уровня – это линия, во всех точках которой значения функции равны.

Авторами статьи [28] было показано, что ишемическая болезнь сердца может быть обнаружена при анализе изображений, полученных на основе биспектра сигнала. Метод описанный в вышеуказанной статье состоял в выделении площади региона внутри линий уровня, при оценке которой можно было сделать вывод о наличии у пациента ишемической болезни сердца. Результаты, которые были получены авторами, позволяют сделать вывод, что изображения биспектра могут использоваться для обнаружения у пациента ишемической болезни сердца.

На Рис. 6 представлен пример изображений, соответствующих биспектральному разложению сигнала для больного ИБС и здорового человека.

Однако следует отметить, что подход, который был реализован в вышеуказанной статье, не учитывает никаких особенностей изображения, кроме площади внутри линий

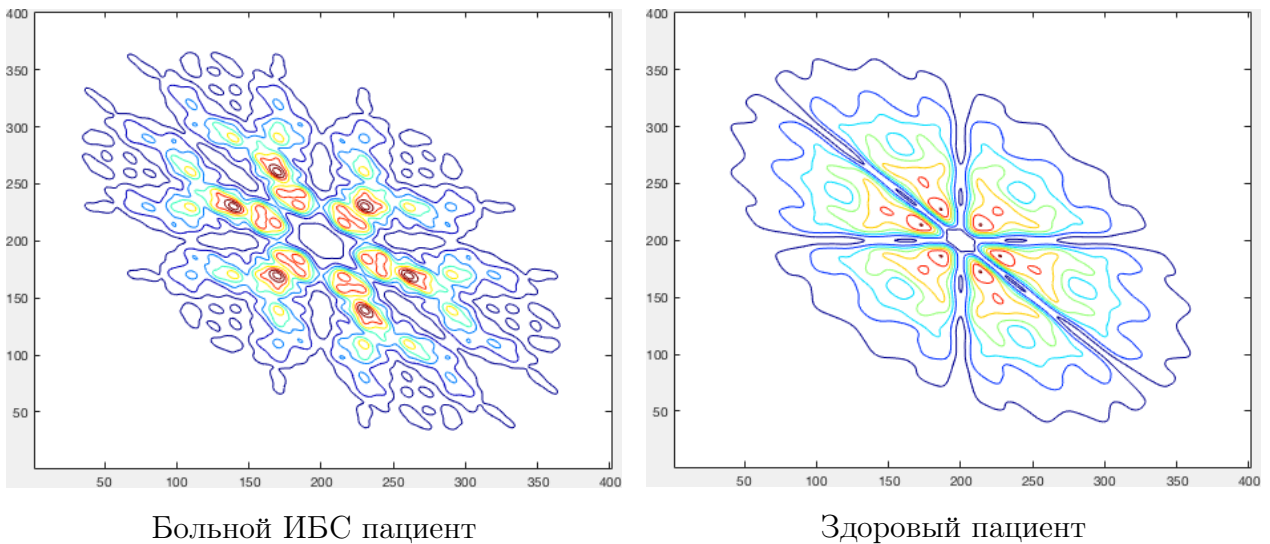


Рис. 6: Изображение, соответствующее биспектральному разложению больного и здорового человека

уровня. Поэтому можно сделать предположение, что проанализировав изображение более сложными методами удастся получить лучшее качество.

В качестве метода анализа данных изображений предлагается использовать нейронные сети, которые на сегодняшний день успешно решают многие задачи, связанные с анализом изображений.

Одним из самых популярных слоев для обработки изображений являются сети, содержащие сверточные слои, которые являются слоями прямого распространения. В сверточном слое каждый нейрон зависит не от всех нейронов предыдущего слоя, а только некоторого набора нейронов, которые находятся в общей окрестности [18]. Такие слои позволяют ускорить процесс обучения и при этом сохранить пространственную информацию. Данный сверточный слой получил свое название из-за использования операции свертки. Суть данной операции заключается в том, что каждый элемент выходного слоя вычисляется как сумма всех элементов поэлементного произведения матрицы ядра свертки и фрагмента изображения, при этом ядро свертки может перемещаться по входному изображению с некоторым заданным смещением по осям.

В данной работе для вычисления биспектра сигнала использовался пакет HOSA [29].

Для классификации полученных изображений, использовалась нейронная сеть с архитектурой, которая описана в таблице 3. Функция активации  $f_{lr}$  – Relu «с уточкой» и она определяется по следующей формуле.

$$f_{lr}(x) = \begin{cases} \alpha x, & \text{если } x < 0 \\ x, & \text{если } x \geq 0 \end{cases},$$

где  $\alpha$  – малая константа.

Входной слой
размер слоя=3x80x80
Сверточный слой
размер фильтра=32x5x5 смещение = (2,2)
Полносвязный слой
число нейронов=30 функция активации = $f_{lr}$
Полносвязный слой
число нейронов=1 функция активации = $f_s$

Таблица 3: Структура нейронной сети

В данной работе нейронная сеть была реализована с помощью библиотек Theano [19] и Lasagne [20], в качестве метода оптимизации использовался алгоритм Adam [18].

## 5 Методы построения ансамблей алгоритмов

Для улучшения качества классификации в поставленной в данной работе задаче было предложено использование ансамблей из реализованных алгоритмов. Ниже описаны методы ансамблирования, которые были использованы в данной работе.

### 5.1 Голосование по большинству

Пусть набор алгоритмов  $A = (A_1, A_2, \dots, A_n)$  выдает вектор ответов  $a = (a_1, a_2, \dots, a_n)$  для классификации случая  $i$ , тогда результирующий ответ  $a$  будет равен

$$a = \text{mode}(a_1, a_2, \dots, a_n),$$

где  $\text{mode}$  — статистика, равная наиболее часто встречающемуся элементу в выборке. В случае если таких элементов несколько, то выбирается один из них случайным образом.

### 5.2 EM-алгоритм

Данный алгоритм был предложен авторами статьи [30] для агрегации данных, полученных от разных людей об одном и том же событии с целью получить истинный

результат. Так как целью построения ансамбля алгоритмов является агрегация ответов различных алгоритмов для получения истинного результата, то данный алгоритм можно использовать для построения ансамбля.

Ниже представлено описание алгоритма для двухклассовой классификации.

$N$  – размер доступных данных,

$n_{il}^k$  – поставил ли  $k$ -ый алгоритм ответ  $l$  ( $l \in \{1, 2\}$ ) на данные  $i$  ( $i = 1 \dots N$ ),

$\pi_{jl}^k$  ( $j \in \{1, 2\}$ ) – вероятность, что  $k$ -ый алгоритм поставит ответ  $j$ , когда истинным ответом является  $l$ .

$T_{ij} = 1$ , если истинным ответом для данных  $i$  является  $j$ , в противном случае, он равняется 0.

$p_j$  – вероятность класса  $j$  в выборке.

- Шаг 1: Инициализируем матрицы  $\pi$  идеальным случаем.  $T$  инициализируем значением голосования по большинству.
- Шаг 2: Пересчитываем значения матриц  $\pi$  и  $p_j$ :

$$\pi_{jl}^k = \frac{\sum_i T_{ij} n_{il}^k}{\sum_l \sum_i T_{ij} n_{il}^k} \quad p_j = \frac{\sum_i T_{ij}}{I}$$

- Шаг 3. Пересчитываем  $T_{ij}$ :

$$T_{ij} = \frac{p_j \prod_{k=1}^K \prod_{l=1}^2 (\pi_{jl}^k)^{n_{il}^k}}{\sum_{q=1}^2 p_q \prod_{k=1}^K \prod_{l=1}^2 (\pi_{ql}^k)^{n_{il}^k}}$$

Повторяем шаги 2 и 3 до тех пор, пока матрицы  $\pi$  не перестанут изменяться.

В конце работы данного алгоритма получим в матрице  $T$  вероятности принадлежности данных каждому классу. В качестве ответа берется тот класс, вероятность принадлежности которому наибольшая.

## 6 Критерии качества, используемые для сравнения реализованных алгоритмов

### 6.1 Качество по пациентам

Данный критерий качества позволяет оценить то, на сколько хорошо алгоритм определяет болезнь человека по любой из его кардиограмм. Кроме того, этот критерий качества не зависит от числа кардиограмм для каждого пациента. Он определяется следующим образом:

$$\sum_{i=1}^N \frac{\sum_{j=1}^{n_i} [t_{ij} == p_{ij}]}{n_i},$$

где  $t_{ij}$  — истинное значение целевой переменной для кардиограммы  $j$  пациента  $i$ ,  $p_{ij}$  — предсказанное значение целевой переменной для кардиограммы  $j$  пациента  $i$ ,  $n_i$  — число кардиограмм у пациента  $i$ ,  $N$  — число пациентов.

## 6.2 F-мера

F-мера — это критерий качества, который не зависит от несбалансированности выборки и позволяет оценить качество модели. Однако следует отметить, что данный критерий качества имеет существенный недостаток, так как не учитывает, что разные пациенты могут иметь различное число кардиограмм. Если алгоритм правильно(неправильно) определяет результат для пациентов с большим числом кардиограмм, то это может исказить результат в большую (меньшую) сторону.

Для определения F-меры введем следующие величины:

- $T_p$  — число правильных ответов модели на объектах из класса 1.
- $T_n$  — число правильных ответов модели на объектах из класса 0.
- $F_p$  — число неправильных ответов модели на объектах из класса 1.
- $F_n$  — число неправильных ответов модели на объектах из класса 0.

$$P = T_p + F_n,$$

$$\text{Полнота (recall)} = \frac{T_p}{P},$$

$$\text{Точность (precision)} = \frac{T_p}{T_p + F_p},$$

$$\text{F-мера} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}.$$

## 6.3 ROC-AUC

Другим критерием качества, который также не зависит от сбалансированности выборки является критерий качества ROC-AUC. Данный критерий позволяет оценить качество вероятностей выдаваемых классификаторами без их преобразования в бинарные метки.

ROC-AUC равен вероятности того, что для случайно выбранных  $x_1 \in C_1$  и  $x_2 \in C_0$  будет выполнено:  $p(x_1) > p(x_2)$ . Где  $C_1$  и  $C_0$  – класс 1 и класс 0 соответственно, а  $p(x)$  – вероятность принадлежности классу  $C_1$ , присвоенная классификатором.

## 7 Результаты экспериментального исследования

### 7.1 Характеристики программной реализации

Программная реализация<sup>1</sup> выполнена на языке Python 2.7.9, с использованием библиотек: pandas [31], numpy [32], а также библиотек, которые были указаны в описаниях алгоритмов.

### 7.2 Метод оценки алгоритмов и подбора параметров

Качество модели  $\mu$  оценивалось на кроссвалидации по 12 фолдам, которая заключается в следующем: выборка  $X$  длины  $L$  разбивается на 12 непересекающихся блоков, таким образом, чтобы в каждом блоке находились разные пациенты и блоки были почти одинаковой длины  $k_1, \dots, k_{12} : X^L = X_1^{k_1} \cup \dots \cup X_{12}^{k_{12}}, k_1 + \dots + k_{12} = L$ . Каждый блок по очереди становится контрольной подвыборкой, при этом обучение производится по остальным 11 блокам. Критерии качества вычисляются на векторе, полученном в результате объединения ответов на контрольных подвыборках в единый вектор и векторе, содержащем истинные значения целевой переменной.

Кроме оценки модели при помощи кросс-валидации также использовалось оценивание модели на отложенной выборке. Оценивание модели на отложенной выборке позволяет оценивать качество предложенных подходов при уже настроенных параметрах.

При обучении моделей настраивались следующие параметры: LogisticRegression –  $C$  (коэффициент регуляризации), XGBClassifier – `learning_rate` (скорость обучения), `n_estimators` (число деревьев), нейронные сети – `learning_rate` (скорость обучения). Остальные параметры были настроены по умолчанию. Параметры подбирались по сеткам. В качестве итоговых параметров выбирались те, при которых качество модели на

---

<sup>1</sup>[https://github.com/ValeryKi/ECG\\_repository](https://github.com/ValeryKi/ECG_repository)



кросс-валидации является наилучшим, или перестает существенно улучшаться относительно критерия качества по пациентам. В качестве итоговой модели выбиралась та, которая дает наилучшее качество по пациентам.

### 7.3 Полученные результаты

Обозначим реализованные методы следующим образом.

BISPECTRUM – алгоритм, основанный на биспектральном разложении сигнала.

WORDS – алгоритм, основанный на вейвлет-преобразовании сигнала (без использования word2vec признаков).

WORDS\_W2V – алгоритм, основанный на вейвлет-преобразовании сигнала (с использованием word2vec признаков).

R\_WINDOW – алгоритм, основанный на анализе сегментов сигнала вокруг его R-пику.

MIX – алгоритм, основанный на 3-х различных группах признаков.

HRV – алгоритм, основанный на признаках, полученных из HRV сигнала.

MAJORITY – ансамбль алгоритмов, выполненный по методу голосования по большинству.

EM – ансамбль алгоритмов, выполненный по EM-методу.

Обозначим алгоритмы классификации как: *xgb* – *xgboost*, *lr* – логистическая регрессия, *nn* – нейронная сеть.

CV обозначает качество модели на кросс-валидации, а T обозначает качество модели на отложенной тестовой выборке.

Полученные результаты приведены в таблицах 4, 5 и 6.

Метод	Алгоритм Классификации	Качество по пациентам (CV)	Качество по пациентам (T)
BISPECTRUM	nn	0.7207	0.7432
WORDS	lr	0.741	0.7699
WORDS_W2V	lr	0.7501	0.7856
R_WINDOW	nn	0.7602	0.7896
MIX	lr	0.744	0.733
MIX	xgb	0.7632	0.7878
HRV	lr	0.7632	0.744
HRV	xgb	0.7693	0.8029
MAJORITY		0.806	0.7605
EM		0.8108	0.8419

Таблица 4: Результаты экспериментальных исследований

Метод	Алгоритм Классификации	ROC-AUC (CV)	ROC-AUC (T)
BISPECTRUM	nn	0.7418	0.724
WORDS	lr	0.8	0.9075
WORDS_W2V	lr	0.8	0.9087
R_WINDOW	nn	0.7988	0.9444
MIX	lr	0.8228	0.9234
MIX	xgb	0.8042	0.9479
HRV	lr	0.6319	0.7506
HRV	xgb	0.744	0.877
MAJORITY			
EM		0.8738	0.9737

Таблица 5: Результаты экспериментальных исследований

Метод	Алгоритм Классификации	F-мера (CV)	F-мера (T)
BISPECTRUM	nn	0.7244	0.7979
WORDS	lr	0.6967	0.7708
WORDS_W2V	lr	0.6990	0.7776
R_WINDOW	nn	0.703	0.789
MIX	lr	0.7057	0.7811
MIX	xgb	0.70256	0.7878
HRV	lr	0.592	0.7490
HRV	xgb	0.6620	0.8029
MAJORITY		0.7784	0.8807
EM		0.7784	0.9024

Таблица 6: Результаты экспериментальных исследований.

Из таблиц 4, 5 и 6 следует, что качество на кросс-валидации по всем критериям качества ниже, чем качество на отложенной тестовой выборке. Данное явление может быть объяснено следующими причинами. Во-первых, на кросс-валидации использовалось меньше данных для обучения, а обучение на полной выборке позволяет добиться лучшего качества. Во-вторых, известно, что первая и вторая выборка содержат разные кардиограммы, однако существует вероятность, что некоторые пациенты из обучающей выборки встречались и в тестовой, а следовательно содержат похожие кардиограммы, в то время как на кросс-валидации одни и те же пациенты не попадали одновременно в обучение и в контроль. Учитывая последний факт, следует отметить, что важно оценивать качество не только на отложенной тестовой выборке, но и на кросс-валидации.

Как видно из Таблицы 4, алгоритм, который основан на биспектральном разложении сигнала, показывает худшее качество по пациентам на кросс-валидации. На отложенной тестовой выборке качество не является худшим, и сопоставимо с качеством по остальным алгоритмам. Заниженное качество по пациентам может быть связано с тем, что обучающая выборка слишком мала для нейронной сети и для того, чтобы получить более высокое качество необходима выборка большего размера. Однако, следует заметить, что качество не слишком низкое, что позволяет сделать вывод, что анализ изображений, полученных на основе биспектрального разложения сигнала может использоваться для определения заболеваний по ЭКГ. В будущем этот метод может быть исследован на предмет повышения качества классификации. Стоит отметить, что этот метод показал лучшее качество по F-мере среди одиночных алгоритмов на кросс-валидации, и одно из лучших на тестовой выборке. Данный факт дает повод предположить, что этот метод

верно идентифицирует пациентов с большим количеством кардиограмм.

Из полученных результатов следует, что при добавлении word2vec признаков модель работает лучше практически по всем выбранным критериям качества (на отложенной тестовой выборке модель по критерию ROC-AUC работает также). Однако прирост не слишком существенен. Следует заметить, что методу word2vec как привило необходимо много данных для того, чтобы получить высокое качество. Поэтому, возможно, что если бы объем выборки был больше, то удалось бы получить более высокое качество при добавлении этих признаков.

Модель, основанная на 3-х различных группах признаков показывает лучшее качество по критерию ROC-AUC на кросс-валидации. На отложенной тестовой выборке она также является лучшей. Это может давать предположение, что данная модель может быть использована, если необходимо выделить группу пациентов, у которых вероятность заболевания является наибольшей.

Наилучшее качество по пациентам на кросс-валидации показывает модель основанная на HRV-сигнале. Однако на отложенной тестовой выборке ее качество сильно зависит от используемого алгоритма классификации. По F-мере данная модель показывает худшее качество на кросс-валидации и худшее качество на отложенной тестовой выборке в случае использования логистической регрессии. Стоит отметить, что данный алгоритм показывает лучшее качество в случае использования градиентного бустинга как на кросс-валидации, так и на отложенной тестовой выборке. Данный факт позволяет предположить, что зависимость между целевой переменной и признаками, входящими в эту модель не является линейной. Можно сказать, что в случае использования градиентного бустинга данный алгоритм дает лучшее качество по пациентам среди одиночных алгоритмов.

Проанализировав результаты в таблице для разных критериев качества можно сделать вывод, что алгоритмы работают по-разному и правильно классифицируют разных пациентов. Данный результат позволяет сделать вывод, что применив ансамбль алгоритмов удастся улучшить качество классификации, что и доказывают результаты для ансамблей алгоритмов как и на кросс-валидации, так и на отложенной выборке. С помощью EM-алгоритма качество удалось существенно увеличить по всем исследуемым критериям качества, как на кросс-валидации так и на отложенной выборке. Это позволяет сделать вывод о том, что ансамбль алгоритмов при использовании EM-алгоритма работает значительно лучше отдельных. Однако, при помощи голосования по большинству не удалось повысить качество по пациентам на отложенной тестовой выборке, но повысилась F-мера. На кросс-валидации голосование по большинству увеличило значения обоих критериев качества, поэтому однозначного решения о его неприменимости в

данной задаче принимать не следует. Следует также отметить, что EM-алгоритм позволяет существенно улучшить качество по ROC-AUC по сравнению с индивидуальными алгоритмами.

## 8 Заключение

В ходе данной работы была проделано следующее.

- Реализовано 5 различных методов обнаружения ишемической болезни сердца по ЭКГ.
- Предложена модификация одного из существующих методов, улучшающая качество классификации.
- Предложен новый метод для классификации ишемической болезни сердца.
- Проведено экспериментальное сравнительное исследование реализованных методов.
- Предложены способы ансамблирования реализованных алгоритмов.
- Исследована применимость EM-алгоритма как метода ансамблирования.

Основными результатами данной работы являются следующие выводы.

- word2vec может увеличить качество классификации метода основанного на вейвлет-преобразовании.
- Биспектральное разложение может быть использовано для классификации ишемической болезни сердца.
- Наилучшее качество по ROC-AUC среди одиночных алгоритмов показывает алгоритм, основанный на 3-х различных признаковых пространствах.
- Наилучшее качество по F-мере показывает алгоритм, основанный на биспектральном разложении.
- Наилучшее качество по пациентам показывает алгоритм, основанный на HRV-сигнале, при использовании в качестве классификатора градиентного бустинга.
- EM-алгоритм применим для ансамблирования и в данном случае показывает наилучшее качество классификации по всем выбранным критериям качества.

- При использовании оборудования CardioQvark возможно определить ишемическую болезнь сердца с точностью свыше 0.8 по качеству по пациентам, с точностью свыше 0.77 по F-мере и с точностью свыше 0.87 по ROC-AUC.

Работа была представлена на международной XXIV научной конференции «Ломоносов - 2017» [33].

## Список литературы

- [1] В.В. Горбачев Ишемическая Болезнь Сердца. - Минск: Выш. шк., 2008. - 479 с. - ISBN 978-985-06-1433-9
- [2] Всемирная организация здравоохранения. Центр СМИ. 10 ведущих причин смерти в мире. [Электронный ресурс] // Всемирная организация здравоохранения. URL: <http://www.who.int/mediacentre/factsheets/fs310/ru/> (дата обращения: 01.05.2017).
- [3] Rangayyan R. M. Biomedical signal analysis. – John Wiley & Sons, 2015. – Т. 33.
- [4] Dua S. et al. Novel classification of coronary artery disease using heart rate variability analysis //Journal of Mechanics in Medicine and Biology. – 2012. – Т. 12. – №. 04. – С. 1240017.
- [5] Giri D. et al. Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform //Knowledge-Based Systems. – 2013. – Т. 37. – С. 274-282.
- [6] Jones E., Oliphant T., Peterson P. SciPy: open source scientific tools for Python. – 2014.
- [7] Hamilton P. Open source ECG analysis //Computers in Cardiology, 2002. – IEEE, 2002. – С. 101-104.
- [8] Carlos Carreiras. BioSPPy – Biosignal Processing in Python. [Электронный ресурс] //BioSPPy’s documentation. URL: <http://biosppy.readthedocs.io/en/stable/> (дата обращения: 01.05.2017).
- [9] Rawald T., Sips M., Marwan N. PyRQA—Conducting recurrence quantification analysis on very long time series efficiently //Computers & Geosciences. – 2016.
- [10] Kantelhardt J. W. et al. Detecting long-range correlations with detrended fluctuation analysis //Physica A: Statistical Mechanics and its Applications. – 2001. – Т. 295. – №. 3. – С. 441-454.
- [11] Christopher Schölzel. Nolds. [Электронный ресурс] // Nolds’ documentation. URL: <https://cschoel.github.io/nolds/> (дата обращения: 01.05.2017).
- [12] Pedregosa F. et al. Scikit-learn: Machine learning in Python //Journal of Machine Learning Research. – 2011. – Т. 12. – №. Oct. – С. 2825-2830.

- [13] Chen T., Guestrin C. XGBoost: Reliable large-scale tree boosting system //Proceedings of the 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA. – 2016. – С. 13-17.
- [14] Hjorth B. EEG analysis based on time domain properties //Electroencephalography and clinical neurophysiology. – 1970. – Т. 29. – №. 3. – С. 306-310.
- [15] De Cooman T. et al. Online seizure detection in adults with temporal lobe epilepsy using single-lead ECG //Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European. – IEEE, 2014. – С. 1532-1536.
- [16] Успенский В. М. Информационная функция сердца. Теория и практика диагностики заболеваний внутренних органов методом информационного анализа электрокардиосигналов. М.: Экономика и информатика, 2008. 116 с.
- [17] Ripoll V. J. R. et al. ECG assessment based on neural networks with pretraining //Applied Soft Computing. – 2016. – Т. 49. – С. 399-406.
- [18] Goodfellow I., Bengio Y., Courville A. Deep learning. Book in preparation for MIT Press //URL< <http://www.deeplearningbook.org>. – 2016.
- [19] Team T. T. D. et al. Theano: A Python framework for fast computation of mathematical expressions //arXiv preprint arXiv:1605.02688. – 2016.
- [20] Dieleman S. et al. Lasagne: First release //Zenodo: Geneva, Switzerland. – 2015.
- [21] Wang J. et al. Bag-of-words representation for biomedical time series classification //Biomedical Signal Processing and Control. – 2013. – Т. 8. – №. 6. – С. 634-644.
- [22] Liu C. L. A tutorial of the wavelet transform //NTUEE, Taiwan. – 2010.
- [23] Wasilewski F. Pywavelets: Discrete wavelet transform in python //Available: <http://www.pybytes.com/pywavelets>. – 2010.
- [24] Mikolov T. et al. Efficient estimation of word representations in vector space //arXiv preprint arXiv:1301.3781. – 2013.
- [25] Rehurek R., Sojka P. Software framework for topic modelling with large corpora //In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. – 2010.
- [26] Civera M., Zanotti Fragonara L., Surace C. Using bispectral analysis and neural networks to localise cracks in beam-like structures. – 2016.



- [27] Nikias C. L., Raghuveer M. R. Bispectrum estimation: A digital signal processing framework //Proceedings of the IEEE. – 1987. – Т. 75. – №. 7. – С. 869-891.
- [28] Al-Fahoum A., Al-Fraihat A., Al-Araida A. Detection of cardiac ischaemia using bispectral analysis approach //Journal of medical engineering & technology. – 2014. – Т. 38. – №. 6. – С. 311-316.
- [29] Swami A., Mendel J., Nikias C. L. HOSA toolbox of Matlab 5.1 //MathWorks, Natick, USA. – 1993.
- [30] Dawid A. P., Skene A. M. Maximum likelihood estimation of observer error-rates using the EM algorithm //Applied statistics, 1979, С. 20-28.
- [31] McKinney W. et al. Data structures for statistical computing in python //Proceedings of the 9th Python in Science Conference. – van der Voort S, Millman J, 2010. – Т. 445. – С. 51-56.
- [32] Walt S., Colbert S. C., Varoquaux G. The NumPy array: a structure for efficient numerical computation //Computing in Science & Engineering. – 2011. – Т. 13. – №. 2. – С. 22-30.
- [33] Гурьянова (Кибитова) В.Н. Ансамбль алгоритмов для определения ишемической болезни сердца // Сборник тезисов XXVI международной научной конференции Ломоносов–2017. Издательский отдел факультета вычислительной математики и кибернетики МГУ имени М.В. Ломоносова, 2017. С. 15-17.