Описание программы (Диагностика)

### 1. Общие сведения

Программный модуль «Диагностика» является составной частью аппаратно-программного комплекса «Киберсердце», предназначен для автоматизации постановки диагноза по имеющимся данным о пациенте. Модуль «Диагностика» имеет сложную структуру и состоит из нескольких подмодулей (подробнее – в разделе 3 «Описание логической структуры»):

1. «Сегментация сигнала ЭКГ»;
2. «Синтез признаков»;
3. «Прямые методы диагностики»;
4. «Методы машинного обучения для постановки диагноза».

Данный программный модуль, получая на вход разнородные данные о пациенте, должен автоматически в качестве результата предоставлять предполагаемый диагноз пациента.

### 2. Функциональное назначение

Можно выделить следующие основные задачи, решаемые модулем «Диагностика»:

1. Чтение исходных данных о пациенте (сигнал ЭКГ, результаты других исследований, анамнез, результаты анализов);
2. Сегментация ЭКГ (назначение подмодуля «Сегментация сигнала ЭКГ») – выделение комплексов и волн на всех полученных отведениях сигнала (временные разметки начала, окончания и пиков каждой волны), а также (опционально) определение их морфологий;
3. Выделение признаков (назначение подмодуля «Синтез признаков») – на основе полученной сегментации сигнала ЭКГ, а также дополнительной информации о пациенте, формирование основных характеристик и признаков, используемых в качестве входа для задачи машинного обучения и прямых методов диагностики;
4. На основе сегментированного сигнала и выделенных признаков автоматическая постановка диагноза, в соответствии с описанными в научной литературе методами диагностики (опционально – синтез дополнительных признаков для задачи машинного обучения) - назначение подмодуля «Прямые методы диагностики»;
5. На основе данных о пациенте и всех выделенных признаков, а также результатов работы прямых методов диагностики, автоматическая постановка диагноза с использованием различных методов машинного обучения;
6. Сохранение итогового диагноза и необходимых промежуточных результатов.

Выделяются следующие основные области применения программного модуля «Диагностика»:

1. Использование в медицинских учреждениях для автоматизации процесса постановки диагноза;
2. Использование в качестве тренажера для обучения студентов-медиков. Основные возможные потребители – медицинские учебные заведения и научно-исследовательские учреждения.

### 3. Описание логической структуры

Высокоуровневое описание алгоритма работы программного пакета «Диагностика» выглядит следующим образом:

1. Чтение исходных данных:

* Данные о пациенте (ФИО, пол, возраст, анамнез);
* Информация об обследовании (дата, тип проводимого обследования, результаты);
* Заключение врача (опционально);
* Сигнал ЭКГ (все доступные отведения);
* Параметры сигнала ЭКГ;

1. Сегментация сигнала ЭКГ на основе алгоритмов, использующих дискретное вейвлет-преобразование, сохранение промежуточной информации о выделенных комплексах и волнах.
2. Выделение признаков на основе исходных данных и полученной промежуточной информации о сегментированном сигнале ЭКГ. База признаков формируется с учетом общепринятых характеристик, используемых в методах медицинской диагностики, а также основных статистических характеристик сигнала. Сохранение полученной базы признаков.
3. Постановка диагноза с помощью прямых методов диагностики. Используются деревья принятия решений, построенные на основе медицинской литературы и консультаций с врачами-кардиологами.
4. Предобработка данных с помощью графового подхода. Данные, пригодные для использования в анализе, представляют собой независимые от времени скалярные величины (для каждого пациента). Не совсем понятно каким образом представить эти данные как единую систему. В частности, какие элементы принадлежат системе и какие находятся вне нее? Если принадлежат, то каковы они будут и существует ли между ними связь? Именно эти задачи и решает графовый подход, с помощью которого для каждого пациента строится паренклитическая сеть, которая, в свою очередь, описывается с помощью топологических индексов, что в дальнейшем позволяет провести анализ новых данных и выявить неочевидные закономерности.
5. Постановка диагноза с помощью методов машинного обучения.
6. Сохранение результатов диагностики.

Разработка программного модуля «Диагностика» ведется следующим образом:

1. Для подмодулей «Сегментация сигнала ЭКГ» и «Синтез признаков для машинного обучения и прямых методов диагностики» используются:

* язык программирования Python;
* методы сегментации сигнала, основанные на дискретном вейвлет-преобразовании;
* методы первичной фильтрации сигнала на основе непрерывного вейвлет-преобразования

1. Для подмодуля «Прямые методы диагностики» используются:

* язык программирования Python;
* деревья принятия решений.

1. Для подмодуля «Методы машинного обучения» и, в частности, графового подхода используются:
   * язык программирования Python;
   * линейная регрессия и метод главных компонент для построения эталонных моделей;
   * логистическая регрессия, случайный лес, система опорных векторов, градиентный бустинг для решения задачи классификации.

Программный модуль «Диагностика» будет включать следующие основные подмодули, активно взаимодействующие с другим программным модулем «Кардиобаза» (*Рисунок 1*):

1. Подмодуль «Сегментация сигнала ЭКГ». Данная подсистема является первичным звеном в работе всего модуля «Диагностика», в которой реализованы алгоритмы по выделению временных интервалов, соответствующих различным волнам и комплексам сигнала ЭКГ (QRS, P, T).

* Непосредственными входными данными данного подмодуля являются сигналы всех доступных отведений ЭКГ, а также параметры сигнала, позволяющие провести настройку алгоритмов сегментации.
* Выходными данными модуля является информация о временной разметке начала, конца и пиков основных комплексов сигнала ЭКГ. Эту информацию используют следующие модули в цепочке обработки данных: «Синтез признаков» и «Прямые методы диагностики» посредством интерфейса модуля «Кардиобаза»

1. Подмодуль «Синтез признаков» на основе исходного сигнала ЭКГ, данных о сегментации, исходных данных о пациенте формирует список характеристик, необходимых для постановки диагноза при помощи подмодулей следующего уровня.
2. Подмодуль «Прямые методы диагностики» производит чтение данных о пациенте и выделенных ранее признаков через интерфейс модуля «Кардиобаза». В процессе диагностики прямыми методами выделяются дополнительные признаки для дальнейшего использования в подмодуле «Методы машинного обучения» (например, является ли ритм синусовым, регулярным и т.д.), которые записываются в базу вместе с поставленным диагнозом и кардиологическим заключением.
3. Подмодуль «Методы машинного обучения», одним из методов которого является графовый подход. На основе синтезированных признаков для каждого пациента строится паренклитическая сеть, которая затем описывается топологическими индексами и в удобной для анализа форме подается на вход стандартным алгоритмам машинного обучения.

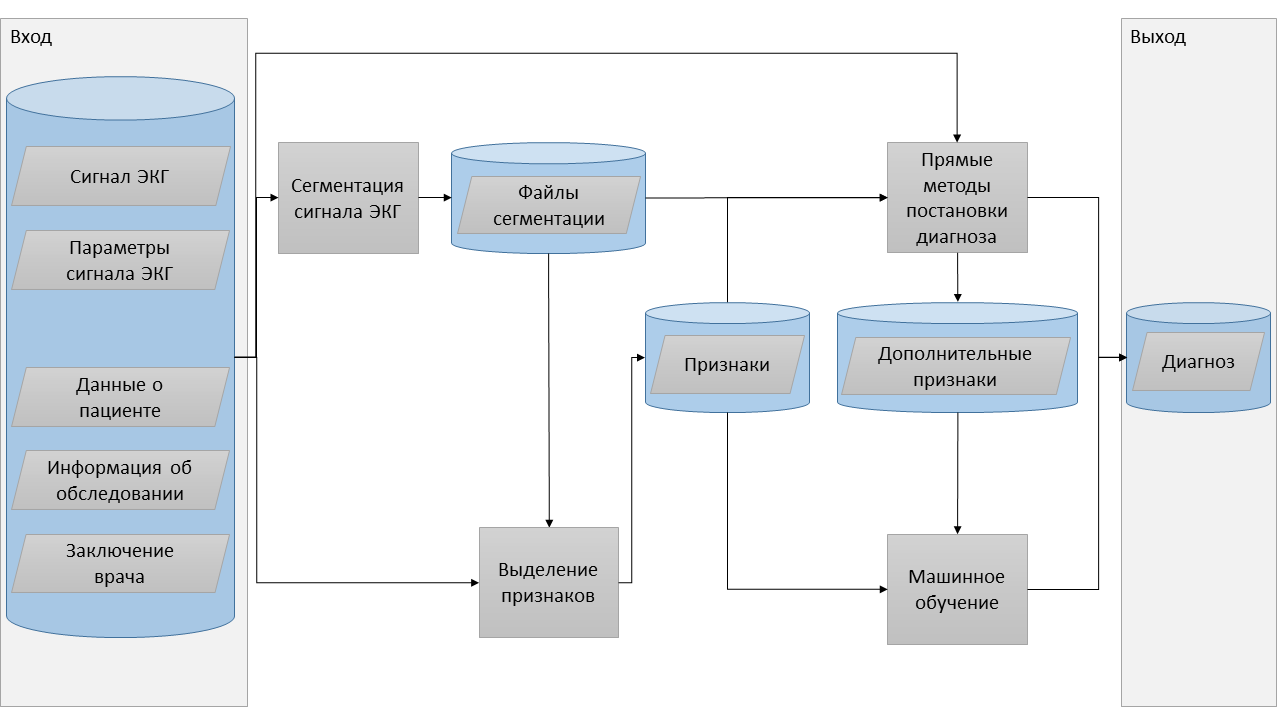


Рисунок 1. Диаграмма подмодулей программного модуля «Диагностика»

### 4. Входные данные

Входные данные программного пакета «Диагностика» предоставляются интерфейсом программного модуля «Кардиобаза».

* Данные о пациенте - ???;
* Информация об обследовании - ???;
* Заключение врача (опционально) - ???;
* Сигнал ЭКГ (все доступные отведения) – одномерные массивы вещественных чисел;
* Параметры сигнала ЭКГ – целочисленная частота сигнала;

### 5. Выходные данные

1. Выходные данные программного пакета «Диагностика» представляют собой отдельный файл для каждого пациента, содержащий итоговый диагноз.

2. Выходными данными подмодуля «Сегментация сигнала ЭКГ» являются отдельные для каждой волны (QRS, P, T) файлы сегментации, представляющие собой многомерные массивы. Первая размерность соответствует количеству найденных волн на исследуемом участке сигнала ЭКГ. Далее, для каждого комплекса записываются: времена начала, окончания, пиков обнаруженной волны, а также тип её морфологии.

3. Выходные данные подмодуля «Выделение признаков» – файлы, содержащие пары элементов: название признака и его значение.

4. Выходные данные подмодуля «Прямые методы диагностики» – файл с диагнозом и кардиологическим заключением, а также дополнительные признаки для машинного обучения в виде файлов, содержащих пары элементов: название признака и его значение.

5. Выходом подмодуля «Методы машинного обучения» является вектор xk, где i-ая компонента вектора показывает вероятность наличия i-ой болезни.

Описание библиотек (Диагностика)

## Диагностика. Парсер для внутреннего формата кардиобазы

### Общие сведения

Представляет из себя библиотеку, работающую как подмодуль модуля “Кардиобаза”, который позволяет преобразовать какой-либо внешний формат во внутренний формат EDF. Программа написана на языке C++ с использованием библиотек edflib, pugixml и csvlib.

### Функциональное назначение

Основное назначение данной библиотеки — преобразование внешних форматов хранения ЭКГ во внутренний формат EDF кардиобазы.

### Описание логической структуры

Библиотека используется как составная часть программ других подсистем модуля “Кардиобаза”.  Цикл работы парсера внутри кардиобазы полностью описывается следующей последовательностью действий.

1. Получение данных ЭКГ и возможных сопутствующих данных в некотором формате.
2. Выделение нужной информации для записи в EDF файл.
3. Запись полученной информации в EDF файл.
4. Запись файла в базу данных.

### Используемые технологии и методы

* Языки программирования С и С++.
* Библиотека pugixml для преобразования xml-файлов.
* Библиотека csvlib для записи массивов данных ЭКГ.
* Библиотека edflib для работы с edf-файлами.

### Входные данные

Входными данными являются сигнал ЭКГ и сопутствующие данные в некотором формате.

### Выходные данные

Выходными данными является EDF файл содержащий сигнал ЭКГ и возможные сопутствующие данные.

## Диагностика. Библиотека работы с модулем “Кардиобаза”

### Общие сведения

Представляет из себя библиотеку, позволяющую взаимодействовать с модулем “Кардиобаза”, который является связующим звеном как между различными модулями проекта “Киберсердце”, так и между подсистемами модуля “Диагностика”. Программы написаны на языке с++ с использованием библиотек oracle, boost и языке python 3.5 с использованием библиотеки \_pickle.

### Функциональное назначение

Основное назначение данного библиотеки — это запись и чтение данных из кардиобазы.

### Описание логической структуры

#### Алгоритм программы

Библиотека используется как составная часть программ других подсистем модуля “Диагностика”.  Цикл работы с кардиобазой полностью описывается следующей последовательностью действий.

1. Подключение к базе
2. Получение необходимой информации о внутренней структуре базы
3. Чтение данных из базы
4. Запись данных в базу
5. Отключение от базы

#### Используемые технологии и методы

* Языки программирования c++ и python 3.5
* Библиотека Oracle для связи с базой данных
* Библиотека boost для создания из с++ кода динамически-линкуемой библиотеки, которую можно использовать в python
* Библиотека сериализации \_pickle