Содержание

Введение

На протяжении всего времени существования человечества проблема возникновения, исследования и лечения различных заболеваний у человека является важной задачей медицинской деятельности. Особенно это касается онкологических заболеваний. На сегодняшний день нет четкой причины, по которой люди заболевают раком, но существует множество способов ранней диагностики таких заболеваний [4].

В данной работе рассматривается использование результатов моделирования, соответствующих методике микроволновой радиотермометрии молочных желез на основе работы специального диагностического комплекса РТМ-01-РЭС [34]. Также рассматриваются популярные алгоритмы классификации данных и библиотеки для языка программирования Python, реализующие данные алгоритмы, и сферы применения данных алгоритмов.

Главной целью работы является разработка программного обеспечения для валидации данных компьютерного моделирования биотканей и возможности диагнорстирования рака молочной железы по температурным данным пациентов с использованием машинного обучения и различных методов классификации. Так как проведение классификации только на температурных данных может не дать хорошего результата, необходимо выявить какой из алгоритмов будет лучше работать с разными размерами опухоли и вариациями остальных параметров.

1 Методы и области применения машинного обучения

1.1 Обзор литературы

Статья Полякова М.В., Хоперскова А.В. «Математическое моделирование пространственного распределения радиационного поля в биоткани: определение яркостной температуры для диагностики», опубликованная в Вестнике Волгоградского государственного университета, посвящена проведению имитационных экспериментов по моделированию динамики температурных и радиационных полей в биотканях молочной железы. В работе вместо традиционно используемых моделей с однородными параметрами используются вычислительные модели максимально приближенные к реалистичной геометрической структуре тканей с неоднородными характеристиками [39].

В статье Веснина С.Г., Седанкина М.К. «Миниатюрные антенны-аппликаторы для микроволновых радиотермометров медицинского назначения», опубликованная в журнале «Биомедицинская радиоэлектроника», описывается анализ миниатюрных антенн-аппликаторов, предназначенных для измерения собственного излучения тканей человека с помощью микроволновых радиотермометров. Приведены простые аппроксимационные формулы для распределения температуры в молочной железе при наличиии злокачественной опухоли [25].

В работе Van Ongeval Ch. «Digital mammography for screening and diagnosis of breast cancer: an overview» обсуждается цифровая телемаммография как новая техника для диагностирования заболеваний молочных желез. Также в данной работе детально рассматривается проблема практической реализации различных систем для визуализации телемаммографической диагностики, высокой стоимости обследования и высокой квалификации специалистов-радиологов [17].

Работа Nisreen I. Yassin, Shaimaa Omran, Enas M. F. El Houby, Enas M. F. El Houby, Hemat Allam «Machine Learning Techniques for Breast Cancer Computer Aided Diagnosis Using Different Image Modalities: A Systematic Review»

посвящена опыту медиков в диагностики и обнаружении рака молочной железы с использованием алгоритмов машинного обучения на основе визуализированных данных обследования пациентов. Целью работы является исследование современного уровня техники в отношении систем компьютерной диагностики и обнаружения рака молочной железы [12].

В статье Левшинского В., Полякова М., Лосева А., Хоперскова А. «Verification and Validation of Computer Models for Diagnosing Breast Cancer Based on Machine Learning for Medical Data Analysis» рассмотрен поход проверки результатов моделирования физических процессов в биотканях с использованием глубокого анализа и машинного обучения. При обучении моделей используются данные измерений температуры пациентов согласно методу радиотермометрии. Так же в работе выделяются на основе набора данных для обучения новые признаки, похожие на те, которые используют медики при обследовании пациентов [9].

В работе Рамсундара Б., Истмана П., Уолтерса П., Панде В. «Глубокое обучение в биологии и медицине» обсуждается применение глубокого обучения в популярных направлениях современных исследований, а особенно в биологии и медицине. Работа содержит описание архитектуры алгоритмов в машинном обучении для применения в задачах данных сферах, а так же некоторые практические примеры по использованию [41].

Статья Jian Ma, Pengchao Shang, Chen Lu, Safa Meraghni «A portable breast cancer detection system based on smartphone with infrared camera» посвящена разработке системы обнаружения рака молочной железы с использованием смартфона с инфракрасной камерой. Для обследования использовался метод инфракрасной термографии и алгоритм классификации k-ближайших соседей. Авторам удалось достигнуть точности определения наличия заболевания больше 98% [6].

В части работ рассмотрен метод микроволновой радиотермометрии и его применение при обследовании рака молочных желез. Так же в некоторых работах рассмотрены способы применения машинного обучения для диагностирования различных заболеваний, в том числе онкологических. Рассмот-

рим далее подробно области применения и основные алгоритмы в машинном обучении.

1.2 Области применения

Использование алгоритмов машинного обучения позволяет решать задачи в различных сферах деятельности человека, таких как недвижимость, сельское хозяйство, экономика, а так же медицина. По данным агенства Frost & Sullivan спрос на разработки, в которых используется машинное обучение в медицине, увеличивается с каждым годом примерно на 40% [22]. Такие разработки могут использоваться как для диагностики заболеваний, так и для биохимических исследований.

Методы машинного обучения активно применяются при медицинском сканировании различных типов, таких как УЗИ или компьютерная томография. Благодаря алгоритмам распознавания образов на изображениях есть возможность анализировать результаты таких исследований и указывать на проблемные участки. Также возможно прогнозирование диагноза пациента по различным его параметрам и результатам исследования. Но программное обеспечение, использующее данные алгоритмы пока не может заменить полностью работу медиков и используется в основном при первичных исследованиях.

При компьютерном моделировании алгоритмы машинного обучения могут использоваться для валидации получившихся данных, или прогнозирования течения каких-либо физических процессов.

1.3 Методы классификации в машинном обучении

Классификация данных состоит из прогнозирования определенного результата на основе уже известных данных. Чтобы предсказать результат, ал-

горитм обрабатывает данные, содержащие набор атрибутов и соответствующий каждому набору результат, обычно называемый атрибутом прогнозирования цели или классом [5]. Формируется модель алгоритма, которая пытается обнаружить отношения между атрибутами, которые позволили бы предсказать результат [7]. Такая процедура называется обучением модели, а набор данных, используемый для этого – тестовой выборкой [10].

Следующим шагом после обучения модели является прогнозирование – процедура определения класса, при которой используется набор данных с неизвестными классами. Такой набор данных, который содержит тот же набор атрибутов, за исключением атрибута прогнозирования, часто называют тестовой выборкой [19].

Алгоритм анализирует входные данные и выдает прогноз. Точность прогноза определяет, насколько «хорош» алгоритм. Например, в медицинской базе данных обучающий набор должен иметь соответствующую информацию о пациенте, записанную ранее, где атрибутом прогноза является наличие или отсутствие у пациента проблем со здоровьем.

Для определения того, какой именно алгоритм использовать для конкретной задачи можно воспользоваться схемой, изображенной на рисунке 1. Исходя из того, что в текущей задаче используется не тестовая информация и имеется 160 примеров, то были выбраны алгоритмы SVM, k-ближайших соседей и наивный байесовский классификатор, описанные ниже.

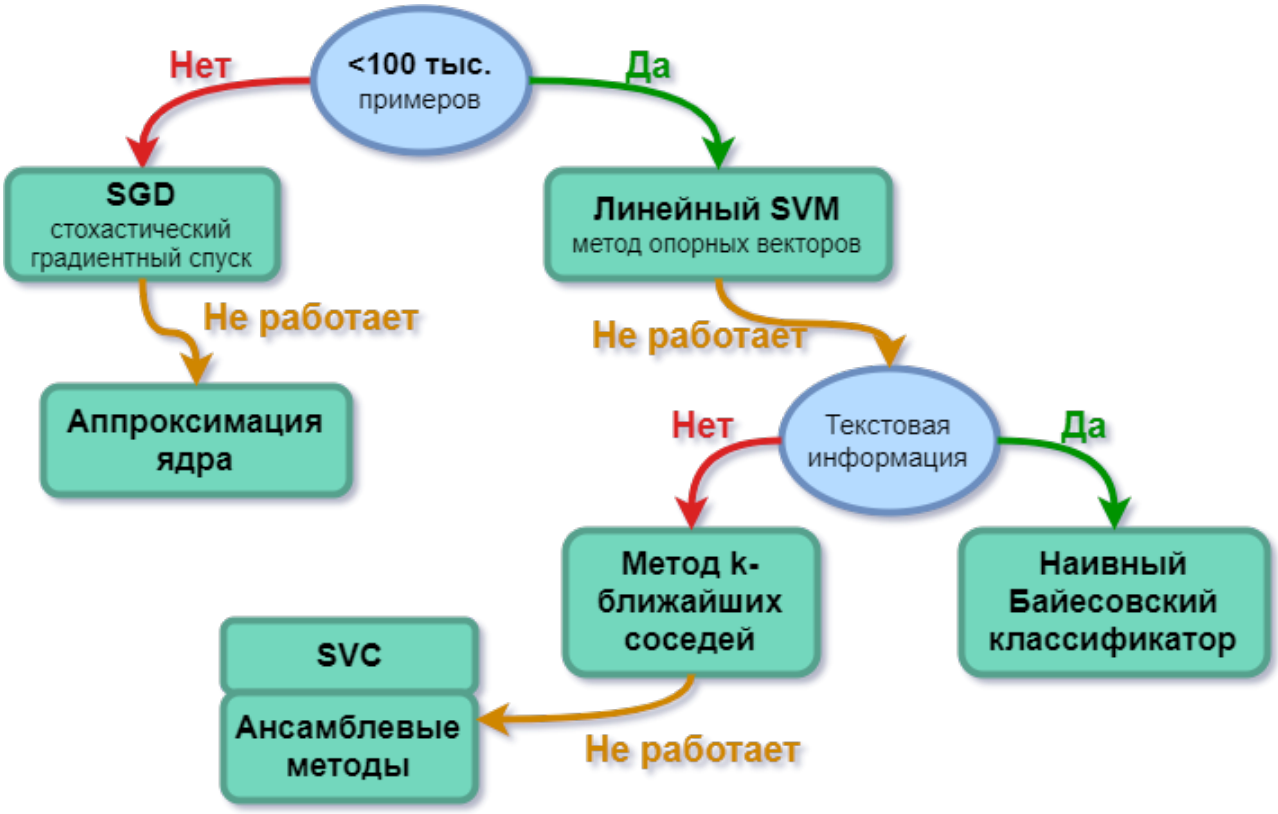

Рисунок 1 – Схема для определения алгоритма классификации для конкретной задачи

1.4 Метод опорных векторов

Метод опорных вектором или SVM – это метод статистической классификации [2]. Он широко используется для задач различного рода и хорошо себя в них показывает [3].

Основной идеей метода является представление атрибутов данных в виде векторов и переход в пространство более высокой размерности, чем получившееся на этапе представления векторами. Затем ищется гиперплоскость с максимальным зазором в пространстве между объектами разных классов [42] [15].

На рисунке 2 показан пример классификации методом SVM. Красной линией выделена как раз та самая гиперплоскость, четко разделяющая объекты разных классов друг от друга.

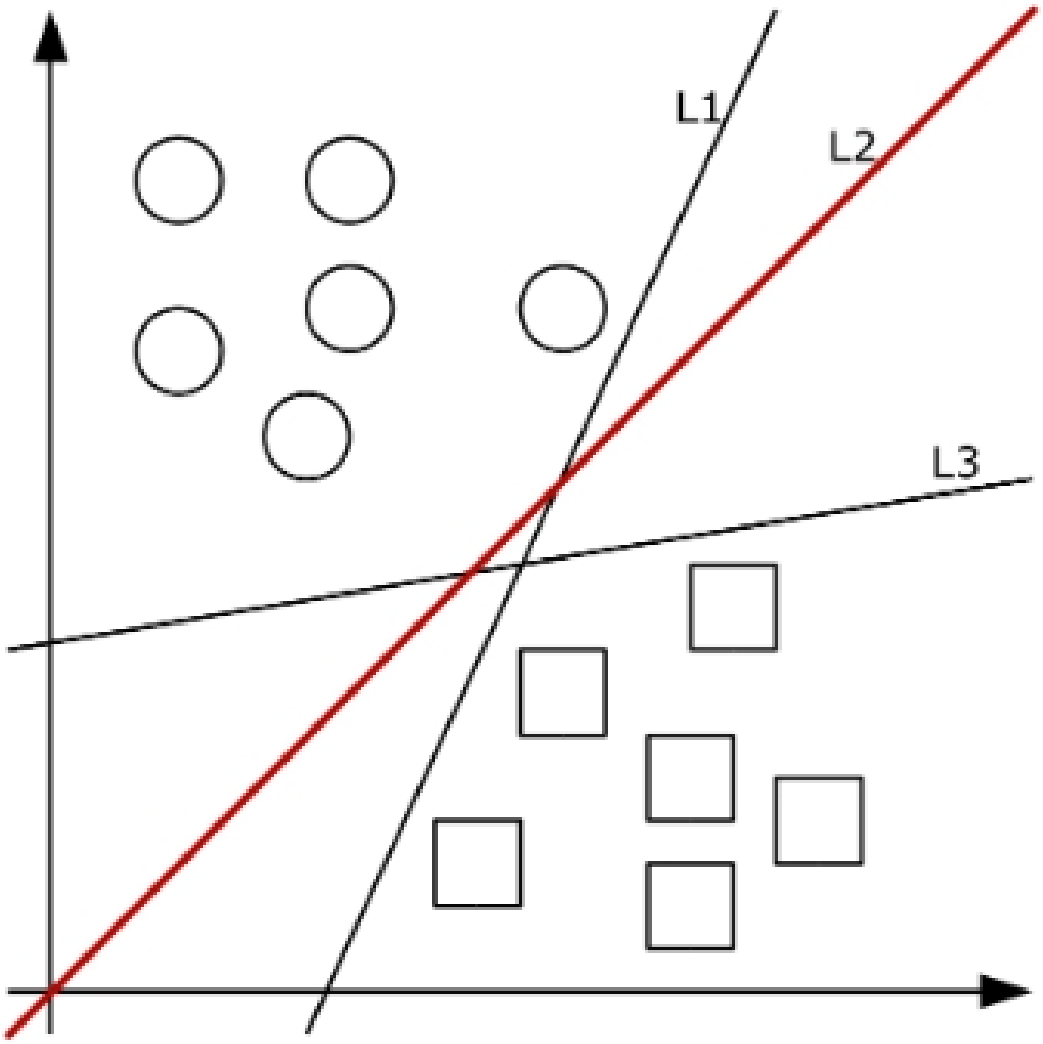

Рисунок 2 – Пример классификации методом SVM

Алгоритм может использоваться с одним из следующих видов ядер [3]:

- Полиномиальное (однородное) $k(x, x') = (x \cdot x')^d$;
- Полиномиальное (неоднородное) $k(x, x') = (x \cdot x' + 1)^d$;
- Радиальная базисная функция $k(x, x') = exp(-\gamma||x - x'||^2)$, для $\gamma > 0$;
- Радиальная базисная функция Гаусса $k(x, x') = exp\left(-\frac{||x-x'||^2}{2\sigma^2}\right)$;
- Сигмоид $k(x, x') = tanh(kx \cdot x' + c)$.

1.5 Метод k-ближайших соседей

Алгоритм k-ближайших соседей является простым статистическим алгоритмом обучения, в котором объект классифицируется своими соседями. При классификации таким методом объект относится к классу, наиболее распространенному среди его k-ближайших соседей [40] [44]. Пример классификации приведен на рисунке 3, где в качестве классифицируемого объекта используется прямоугольник и существует несколько объектов известных классов – белые точки и черные. Замерив расстояние от объекта до его соседей с различными классами и основываясь на методе k-ближайших соседей данный объект будет отнесен к классу черный точек, а не белых.

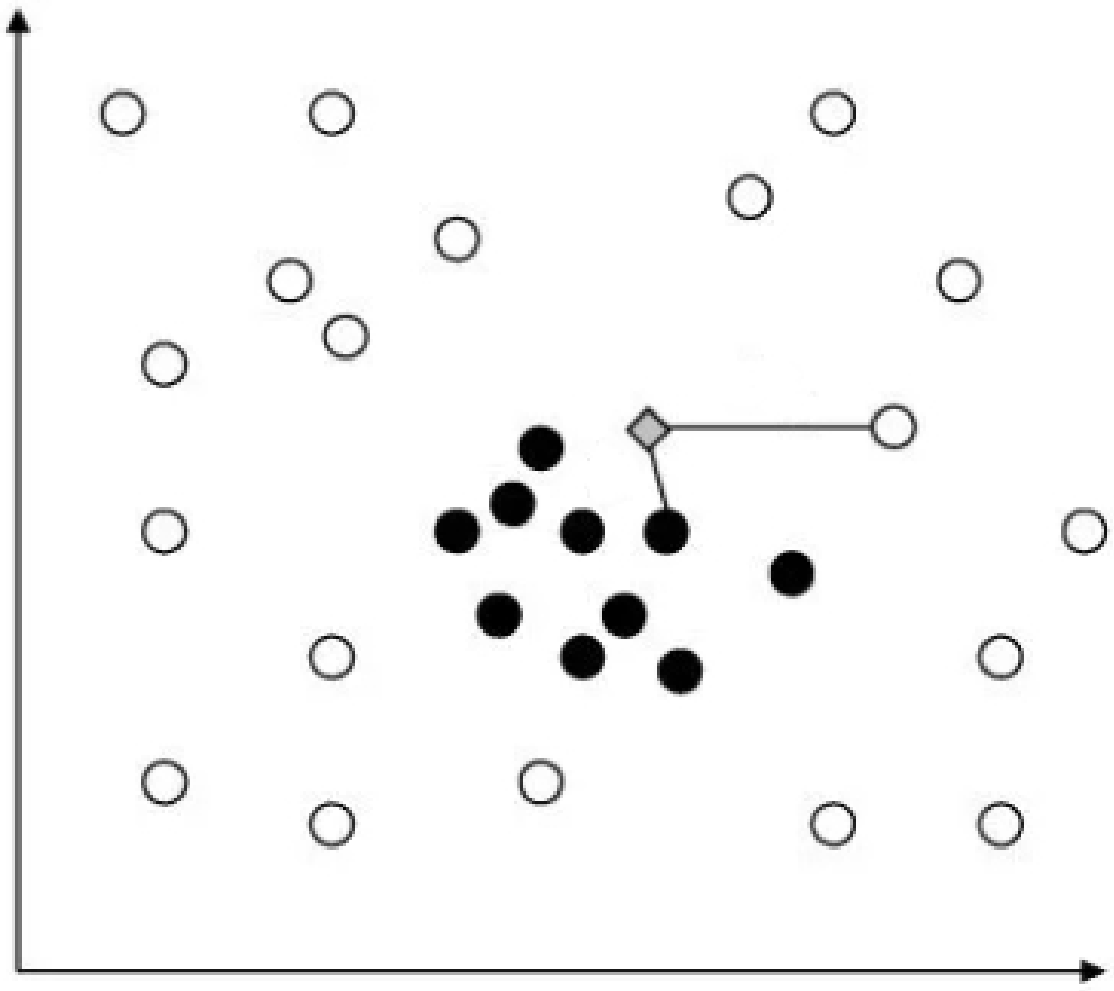

Рисунок 3 – Пример классификации методом k-ближайших соседей

При нахождении атрибутов учитывается значимость атрибутов и часто применяется прием растяжения осей, демонстрируемый в формуле (1). Использование данного приема снижает ошибку классификации.

$$D_E = \sqrt{3(x_A - y_A)^2 + (x_B - y_B)^2}, \tag{1}$$

где x_A, y_A – значения атрибута A в наборе данных, x_B, y_B – значения атрибута B.

Данный алгоритм возможно применять как для данных с маленьким количеством атрибутов, так и с достаточно большим. Важным моментом при работе с алгоритом является определение функции расстояния между значениями. Примером такой функции может быть евклидово расстояние – формула (2).

$$D_E = \sqrt{\sum_{i}^{n}(x_i - y_i)^2}, \tag{2}$$

где x_i, y_i – значения атрибутов в наборе данных.

1.6 Наивный байесовский классификатор

Наивный байесовский классификатор является простым вероятностным классификатором и основывается на применении теоремы Байеса со строгими предположениями о независимости [40] [29]. Хотя наивный байесовский классификатор редко применим к большинству реальных задач, но зачастую в определенных задачах он демонстрирует хорошие результаты и часто конкурирует с более сложными методами, такими как SVM и классификационным деревьями [10]. Классификация данным методом очень зависит от распределения зависимостей атрибутов, а не от самих зависимостей [32].

Вероятностная модель классификатора:

$$p(C|F_1, ..., F_n) = \frac{p(C)p(F_1, ..., F_n|C)}{p(F_1, ..., F_n)}, \tag{3}$$

где C – класс модели, а F_i – классифицируемые модели [10].

Использование формулы (3) при классификации дает минимально значение среднего риска или математичского ожидания ошибки:

$$R(a) = \sum_{y \in Y} \sum_{\varsigma \in Y} \lambda_y P_y P_{x,y} \{a(x) = \varsigma | y\}, \tag{4}$$

где λ_y – цена ошибки при отнесении объекта класса Y к какому-либо другому классу.

1.7 Популярные библиотеки с реализацией методов машинного обучения

На текущий момент существует множество готовых реализаций алгоритмов машинного обучения и не имеет смысла делать то же самое с нуля, если задача не имеет каких-то особенностей, делающих невозможным использование готовых библиотек. Каждая из библиотек, рассматриваемых в работе, хороша в своей области, успешно используется в решении задач и проверена временем. Рассмотрим некоторые из популярных библиотек для языка программирования Python по данным рейтинга на GitHub (рисунок 4)

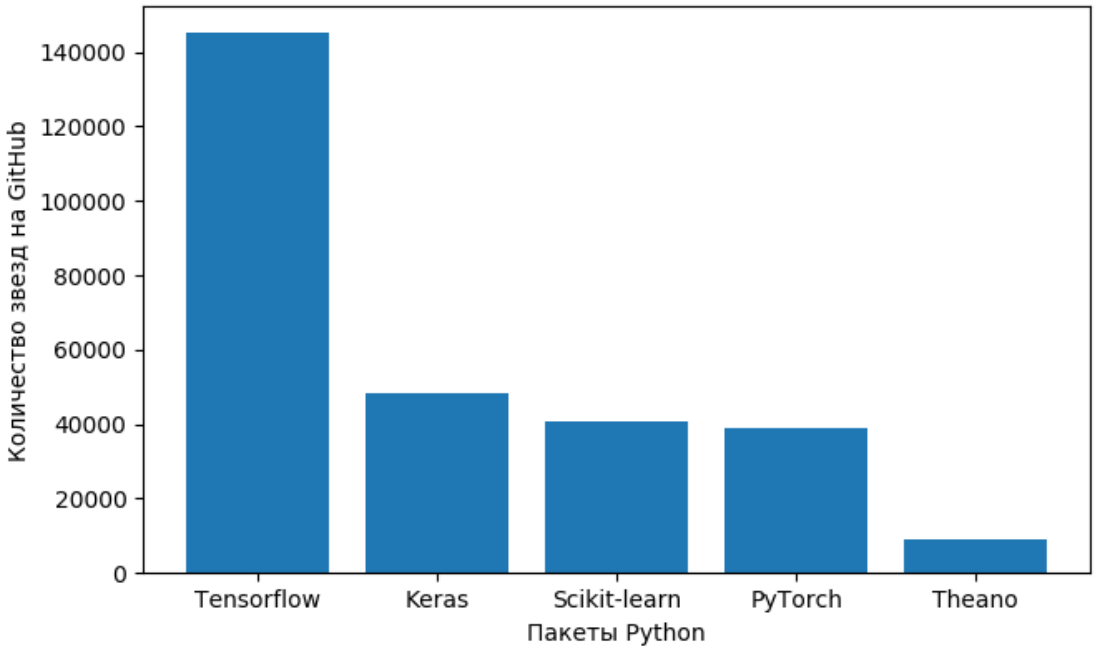

Рисунок 4 – Популярные пакеты Python для машинного обучения по данным рейтинга на GitHub

1.7.1 TensorFlow

Самой популярной и масштабной по применению является библиотека TensorFlow, используемая для глубокого машинного обучения [28]. Библиотека разрабатывается в тесном сотрудничестве с компанией Google и применяется в большинстве их проектов где используется машинное обучение. Библиотека использует систему многоуровневых узлов, которая позволяет вам быстро настраивать, обучать и развертывать искусственные нейронные сети с большими наборами данных.

Библиотека хорошо подходит для широкого семейства техник машинного обучения, а не только для глубокого машинного обучения. Программы с использованием TensorFlow можно компилировать и запускать как на CPU, так и на GPU. Также данная библиотека имеет обширный встроенный функционал логирования, собственный интерактивный визуализатор данных

и логов [36].

1.7.2 Keras

Keras используется для быстрого прототипирования систем с использованием нейронных сетей и машинного обучения. Пакет представляет из себя высокоуровневый API, который работает поверх TensorFlow или Theano. Поддерживает как вычисления на CPU, так и на GPU

1.7.3 Scikit-learn

Scikit-learn – это одна из самых популярных библиотек для языка Python, в которой реализованы основные алгоритмы машинного обучения, такие как классификация различных типов, регрессия и кластеризация данных. Библиотека распространяется свободно и является бесплатной для использования в своих проектах [42].

Данная библиотека создана на основе двух других – NumPy и SciPy, имеющих большое количество готовых реализаций часто используемых математических и статистических функций. Библиотека хорошо подходит для простых и средней сложности задач, а также для людей, которые только начинают свой путь в изучении машинного обучения.

1.7.4 PyTorch

PyTorch – это популяный пакет Python для глубокого машинного обучения, который можно использовать для расширения функционала совместно с такими пакетами как NumPy, SciPy и Cython. Главной функцией PyTorch является возможность вычислений с использованием GPU. Отличается высо-

кой скоростью работы и удобным API-интерфейсом расширения с помощью своей логики, написанной на C или C++.

1.7.5 Theano

Theano – это библиотека, в которой содержится базовый набор инструментов для машинного обучения и конфигурирования нейросетей. Так же у данной библиотеки есть встроенные методы для эффективного вычисления математических выражений, содержащих многомерные массивы [42].

Theano тесно интегрирована с библиотекой NumPy, что дает возможность просто и быстро производить вычисления. Главным преимуществом библиотеки является возможность использования GPU без изменения кода программы, что дает преимущество при выполнении ресуркоемких задач. Также возможно использование динамической генерации кода на языке программирования C [31].

2 Проектирование и разработка программы для обработки данных компьютерного моделирования биотканей

Для имеющихся результатов компьютерного моделирования биотканей необходимо было разработать программное обеспечение, которое позволяло бы определять насколько смоделированные данные соответствуют реальным, а так же для возможности проводить тесты на данных реальных пациентов в будущем.

Для решения данной задачи хорошо подходят методы машинного обучения, а в частности задача классификации, т.к. имея некоторый набор данных, можно обучить модель и настроить параметры для более точной работы в дальнейшем. При обучении модели с помощью метрики точности определения класса можно будет судить о качестве обучения.

2.1 Формат входных данных для программы

В работе использовались данные компьютерного моделирования яркостной температуры молочных желез больных и здоровых пациентов. Данные были представлены в виде девяти значений температуры на поверхности кожи и девяти значений внутренней температуры, согласно методике обследования методом радиотермометрии [4] [1]. Схема расположения точек при замере температур представлена на рисунке 5 [24]. Отдельным атрибутом является класс модели. Для здоровых моделей значения класса было равно нулю, а для больных – единице. Исходя из количества классов, классификацию в данной работе можно считать бинарной.

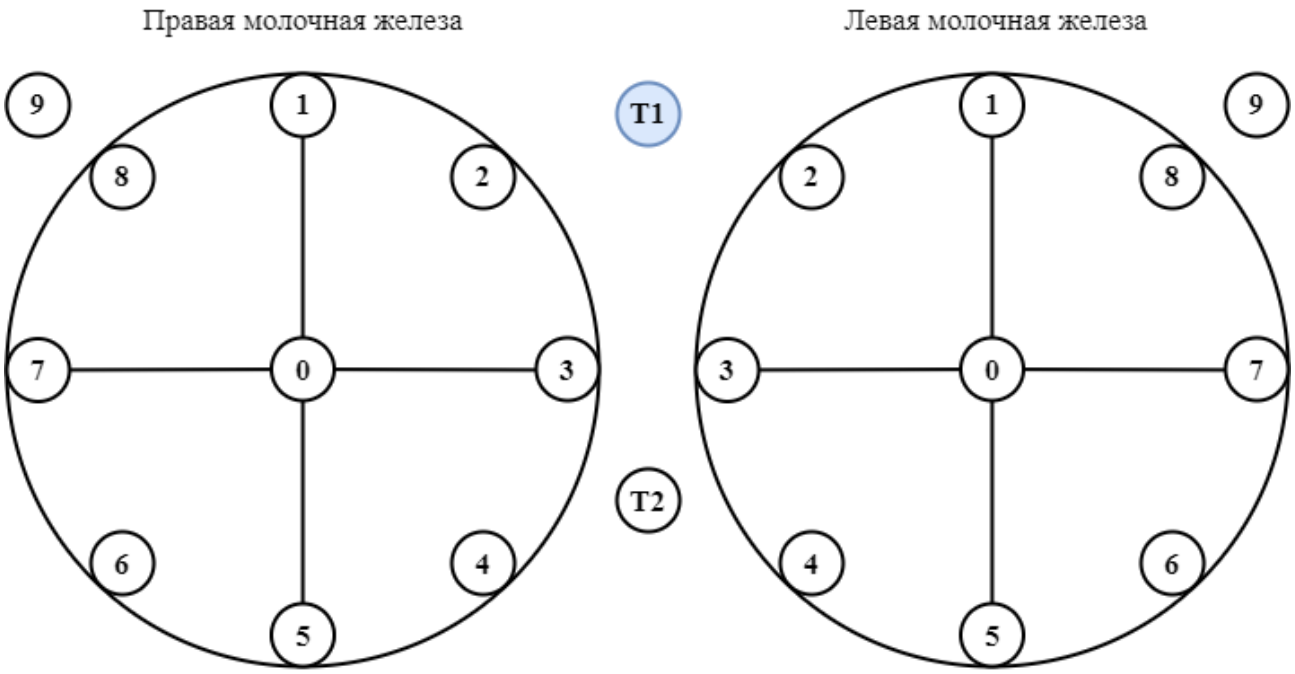

Рисунок 5 – Схема расположения точек при замере температур методом РТМ

Для исследования были взяты температурные данные моделей с радиусом опухоли 0.5 см и 0.75 см. Данные были представлены в виде CSV-файла, в котором находилось по 160 моделей для каждого размера опухоли соответственно (всего 320 моделей) (рисунок 6).

0ртм	1ртм	2ртм	3ртм	4ртм	5ртм	6ртм	7ртм	8ртм	0ик	1ик	2ик	3ик	4ик	5ик	6ик	7ик	8ик	target	point
35,5	35,5	35,6	35,4	34,9	35,2	35,1	35,3	35,6	32,3	33,9	33	33,4	33,4	33,1	32,5	33,2	34,4	0	10
34,4	33,5	33,4	33,5	34,2	34,1	34,4	34,1	34,6	32,2	32	31,9	31,4	32,6	32,8	31,9	32	32,8	0	10
35,7	34,6	34,5	34,4	34,9	34,2	34,2	33,9	34,5	33,8	33,1	32,6	33,2	33,4	32,2	31,8	31,6	32,6	1	0
33,3	32,7	32,9	33,5	34,1	33,5	32,8	33,4	33,8	31,6	32,3	31,7	32,4	32,7	31	30,6	31,4	32,6	1	3
33,6	32	33,3	32,5	32,6	33,5	33	33	32,6	31,8	32,2	33,4	30,9	31,9	31,5	31	31,4	30,8	1	5
34,4	34,4	34,4	34,6	34,3	33,5	33,3	33,5	34,2	33,1	33,2	33	33,3	32,5	32,3	31,8	31,8	32,5	1	4
34,2	33,5	33,8	33,5	33,8	34,1	33,5	33	33,3	32,1	31,5	31,8	30,6	31,1	31,3	30,9	30,3	30,4	1	0
33,1	32,9	33,3	33	33,7	33,3	32,5	32	32,4	31,4	31,7	31,7	31,5	32,1	31,5	31,1	30,5	30,1	1	4
35,5	34,8	34,7	34,7	35,1	35	34,6	34,7	34,8	33,6	33,3	33,3	33,9	33,9	33,4	33,1	32,9	33,4	1	0
33,6	32	32,2	32,7	33	32,9	31,9	32,5	33,5	31,6	30,6	30,7	31,4	31,6	31,4	30,3	31,9	31,7	1	8
36	35,4	35,6	35,7	35,7	35,4	34,9	34,8	35,1	34,3	34,3	34,7	34,4	34,5	33,8	33,1	32,5	33,2	1	2
35,5	34,1	34,5	35,2	34,6	34,9	34,4	33,7	35,4	34,2	33,3	33,6	34	33,8	33,6	32,7	32,2	34,7	1	8
31,5	31,8	32,4	33,6	33,2	33	32	31,7	31,7	29,6	29,1	29,4	30,9	31	30,4	29,7	29,3	28,8	1	3
35	33,7	34	34,4	34	33,9	34,1	33,3	33,6	33,2	32,5	32,3	32,6	32,4	32,5	31,8	31,5	32,2	1	0
31,8	31,9	34,1	32,3	33,3	32,9	32,6	32,8	31,8	29,9	29,9	32,3	30	31,6	30,5	29,9	29,4	30,5	1	2
33,5	33,1	32,8	33,6	33,5	34,2	33,1	33,3	33,3	31,2	29,9	30,4	31	31,3	31,2	30,1	30,2	29,9	1	4

Рисунок 6 – Пример данных температурных данных компьютерного моделирования, где в столбце «target» здоровые — «0», больные — «1»

Одна половина моделей состоит из здоровых пациентов, а другая из больных. В столбце point находятся данные о том, к какой точке ближе всего располагается опухоль. Эти данные можно использовать для локализации

опухоли в дальнейшем. На подготовительном этапе данные были разбиты на обучающую и тестовую выборки. По умолчанию тестовая выборка бралась как 25% от всех данных.

2.2 Проектирование структуры программы и интерфейса

В разрабатываемой программе должна быть возможность определить данные для обучения и классификации, обучить модель с заданными параметрами и методами. Для каждого метода необходимо после обучения и классификации тестовой выборки расчитать точность определения класса и показатели информативности диагностики. Одними из интересных для текущей задачи показателями являются чувствительность (5) и специфичность (6).

$$Se = \frac{TP}{TP + FN}, \tag{5}$$

где TP – количество истинно положительных результатов, FN – количество ложноотрицательных результатов.

$$Sp = \frac{TN}{TN + FP}, \tag{6}$$

где TN – количество истинно отрицательных результатов, FP – количество ложноположительных результатов.

После обучения модели необходимо дать пользователю возможность протестировать ее на данных пациента и показать результат в виде класса и точки с опухолью. Для контроля хода обучения и последующей корректировки параметров обучения будет полезным отображение статистических данных о выборках и графики с точностью классификации.

После определения того, что может сделать пользователь в программе и что он увидит в результате, была разработана диаграмма деятельности (рисунок 7). Данная диаграмма будет полезной как при разработке, так и при тестировании, т.к. содержит последовательную схему действий пользователя.

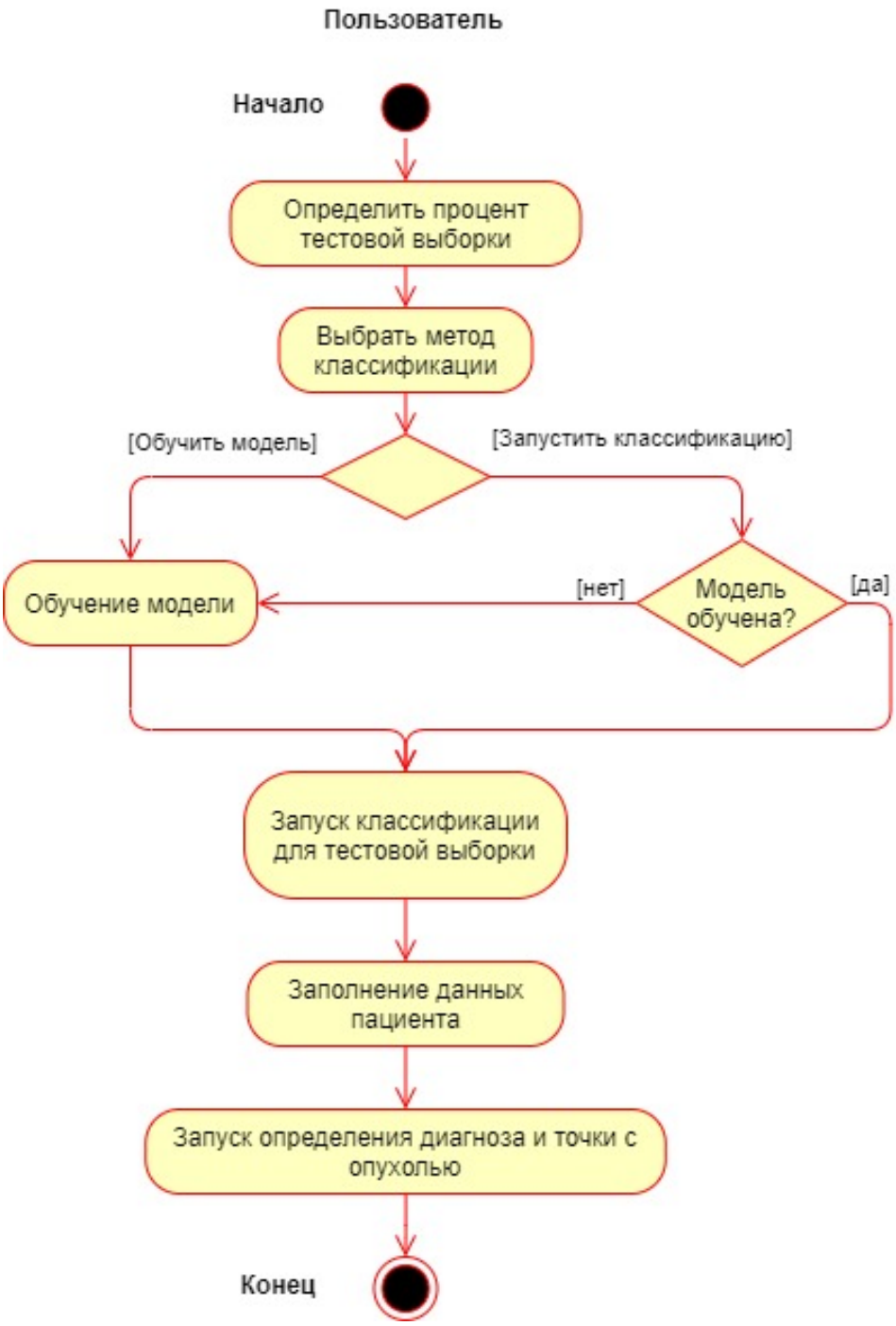

Рисунок 7 – Диаграмма деятельности для программы

После определения возможных действий пользователя был разработан макет интерфейса программы (рисунок 7). При разработке макеты были учтены требования по возможностям настройки и контроля процесса обучения моделей.

Рисунок 8 – Макет интерфейса программы

Интерфейс программы разделен на две части: слева находятся элементы управления для настройки параметров обучения, справа – результаты обучения и статистические данные по выборкам. Для каждого метода классификации будет использован одинаковый набор полей с настройками и одинаковый набор графиков. В блоке с графиками результаты обучения представлены как круговые диаграммы с точностью. Так же имеются графики со статистическими данными – частотным распределением опухолей по точкам и распределением температуры по точкам.

2.3 Выбор технологий и архитектуры

Перед началом разработки программы встал вопрос о том, с помощью каких технологий она будет реализована. Если в самом начале выбрать неправильные инструменты для разработки, то в дальнейшем это может сильно

усложнить поддержку программного обеспечения.

В качестве языка программирования был выбран Python и библиотека Scikit-learn, т.к. для них есть множество примеров использования под текущую задачу и обучающих материалов.

Для пользователей программы было бы удобно не иметь копию данных с результатами моделирования, т.к. файл с этими данными может быть достаточно большого размера. Если данных будет слишком много, то модель будет гораздо дольше на таком наборе данных. Исходя из этого было принято решение использовать клиент-серверную архитектуру при разработке. Взаимодействие клиента и сервера можно условно разделить на две части:

- Загрузка данных, обучение и классификация тестовой выборки;
- Определение диагноза пациента и локализация опухоли.

На рисунках 9 и 10 показаны диаграммы последовательности, описывающие взаимодействие клиента и сервера для каждой из частей. Для общения клиента и сервера был выбран протокол HTTP из-за большой поддержки во многих языках программирования, библиотеках и фреймворках.

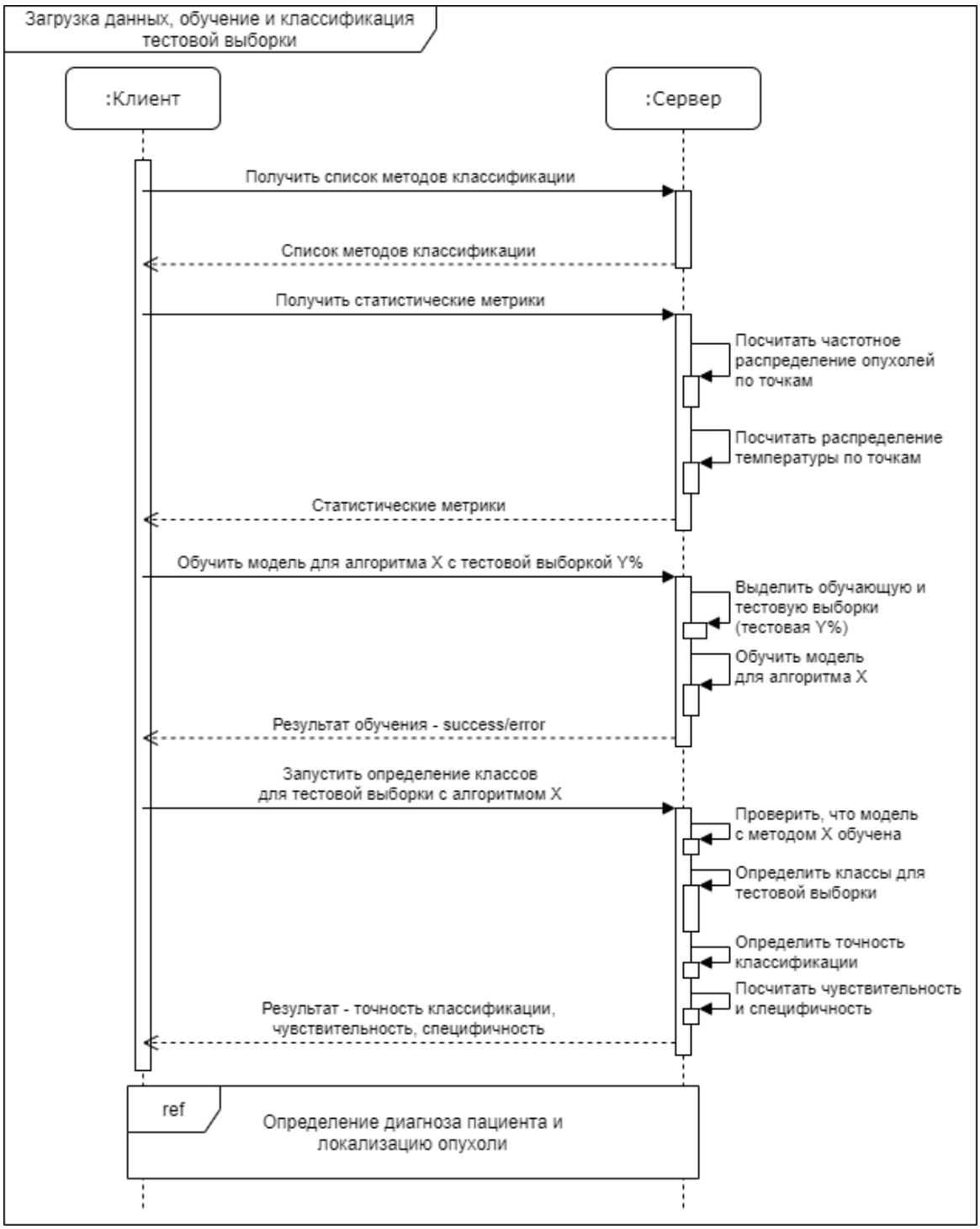

Рисунок 9 – Диаграмма последовательности для этапа загрузки данных, обучения и классификации тестовой выборки

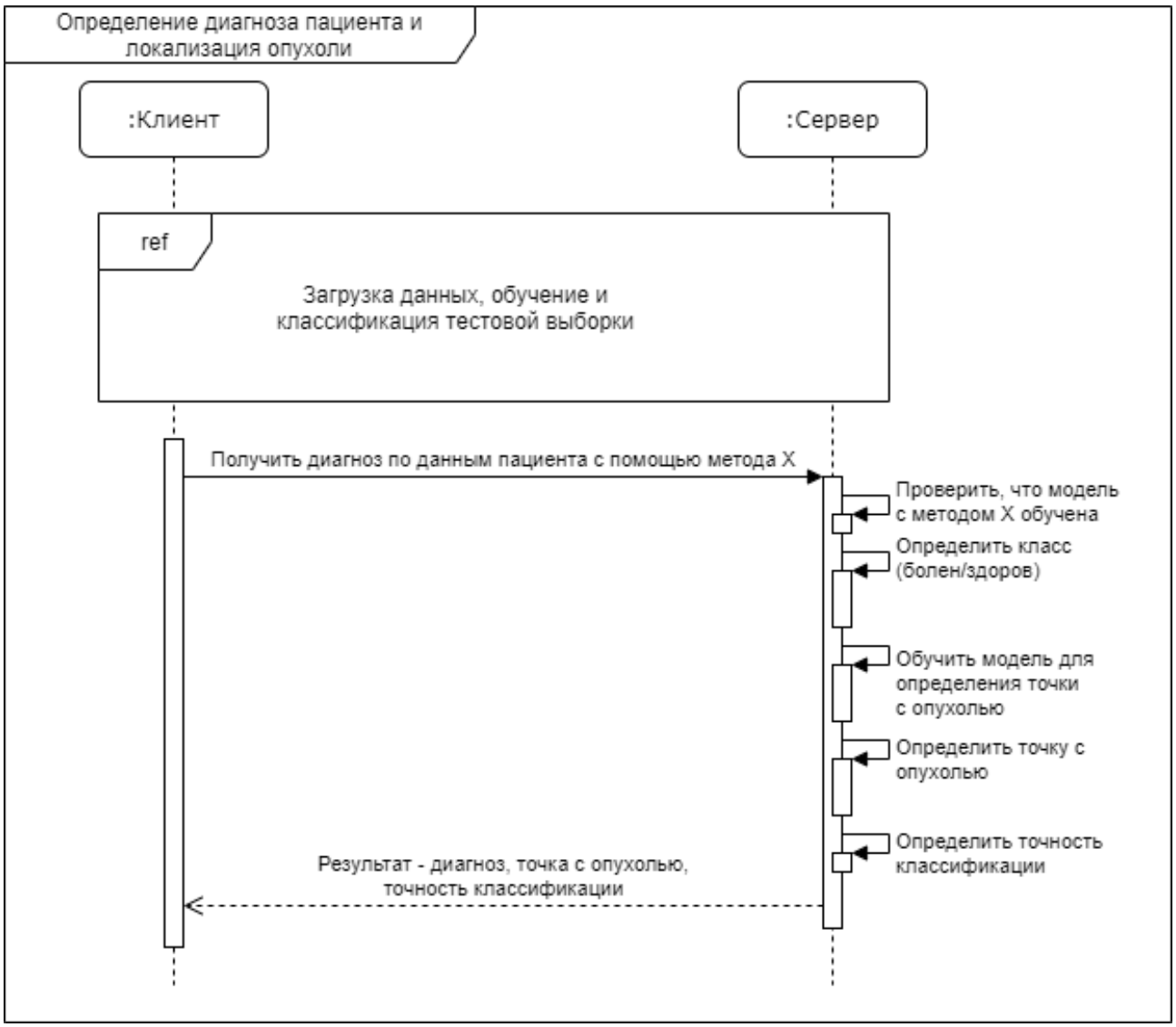

Рисунок 10 – Диаграмма последовательности для этапа определения диагноза пациента и локализации опухоли

В силу простоты реализации и возможности быстрого прототипирования API-интерфейса для бэкенд-части был выбран веб-фреймворк Flask. Flask имеет множество дополнительных библиотек для расширения функционала, а так же подробную документацию.

Для разработки интерфейса рассматривались такие библиотеки для языка Python как Kivy, Tkinter и PyQt. Каждая из них имеет большие возможности для визуализации данных и реализации различных элементов интерфейса. Так же был рассмотрен вариант реализации веб-интерфейса, который и был выбран в дальнейшем из-за возможности его использования на различных типах устройств. Вторым плюсом веб-интерфейса является простота

обновления программного обеспечения в будущем, т.к. такой вариант реализации не требует от пользователя загрузки и установки программы к себе на устройство. Для разработки современного и быстро работающего без перезагрузки страницы интерфейса был выбран язык программирования JavaScript и фреймворк VueJS. Приложение на VueJS состоит из отдельных компонентов, каждый из которых имеет свое состояние и свойства. Такой подход позволяет переиспользовать компоненты и удобно настраивать взаимодействие между ними.

2.4 Разработка программы

На рисунке 11 изображена структура проекта. В отдельных директориях хранятся стили, JS-файлы, шаблоны и тесты.

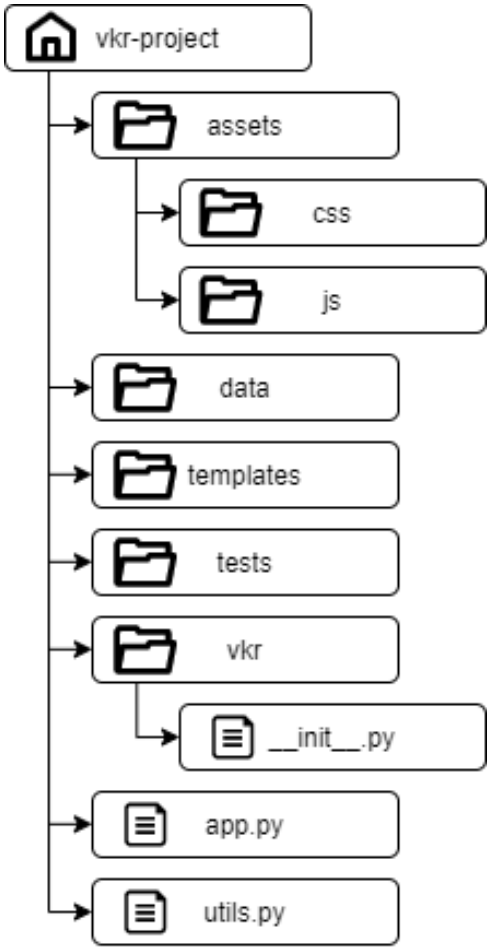

Рисунок 11 – Схема структуры проекта

Разработка программы в рамках данной работы была разделена на этапы, каждый из которых детально описан далее.

2.4.1 Реализация API-интерфейса

Бэкенд-часть представляет собой веб-приложение на Flask. Приложение состоит из одного файла app.py, в котором инициализируется объект приложения с определенными настройками. При запуске файла app.py приложение запускается и становится доступным по адресу http://127.0.0.1:5000/. Также в этом файле описываются все маршруты приложения. В рамках данной работы были реализованы следующие маршруты:

- / – главная страница приложения, с которой и работает пользователь;
- /methods/ – API-метод для получения доступных методов классификации;
- /static_metrics/ – API-метод для получения статистических метрик, рассчитанных для текущей выборки данных;
- /upload_data/ – API-метод для загрузки CSV-файла с температурными данными;
- /train/ – API-метод для обучения модели с выбранным алгоритмом классификации;
- /predict/ – API-метод для запуска классификации для тестовой выборки;
- /diagnose/ – API-метод для получения диагноза пациента по его температурным данным.

Для удобства разработки был создан модуль vkr, где был размещен весь код, связанный с классификацией и обработкой данных.

Чтение файла с данными и инициализация методов для обучения происходит при старте приложения, а так же после загрузки пользователем нового файла.

Для разделения данных на обучающую и тестовую выборки с опре-

деленным соотношением была использована функция train_test_split() из библиотеки Scikit-learn. Во время ее вызова ей необходимо передать массив данных, размер тестовой выборки в процентном соотношении и нужно ли перемешивать данные при разбиении.

Чтобы не обучать модель каждый раз, при первом обучении она сохраняется в файл в бинарном виде. При последующих обращениях к объекту модели, сначала вызывается функция, которая проверяет наличие такого файла. Если файла нет – то модель обучается и он создается. Если файл есть – то данные берутся из него. Это сделано с помощью пакета Pickle. Он позволяет экспортировать в файл и импортировать из файла переменные любых типов.

При определении диагноза пациента происходит еще и определение точки, в которой находится опухоль. При определении точки создается новая модель с алгоритмом многослойной классификации с Перспептроном и обучается на данных только больных пациентов. В качестве классов используются данные из столбца point. Результатом классификации является номер точки. После определения точки расчитывается точность классификации.

Список доступных методов хранится на стороне бэкенда, что позволяет создать универсальный интерфейс на фронтенде для работы с ними.

2.4.2 Разработка фронтенд-части и связь с API

Для разработки фронтенд-части был использован подход работы с VueJS как с библиотекой, подключаемой на странице, т.е. не использовалась сборка Webpack или Vue CLI. Этот выбор связан с небольшим количеством компонентов. Если в будущем при какой-либо масштабной доработке количество компонентов начнет стремительно увеличиваться, то переход на вариант со сборкой не будет большой проблемой, ведь все Vue-компоненты были вынесены в отдельный JS-файл и имеют схожую структуру. В зависимости от того, в каком режиме запущен проект (для разработки или как «боевой») подклю-

чается либо версия VueJS «dev-версия» для разработки, либо «production-версия». Главное отличие этих версий в том, что в «production-версии» отсутствуют инструменты для отладки и минифицирован код, чтобы файл с библиотекой занимал меньше места.

На начальном этапе были выделены Vue-компоненты (рисунок 12). Данные о методах классификации, текущем выбранном методе и статистические данные будут храниться в главном компоненте с названием App.

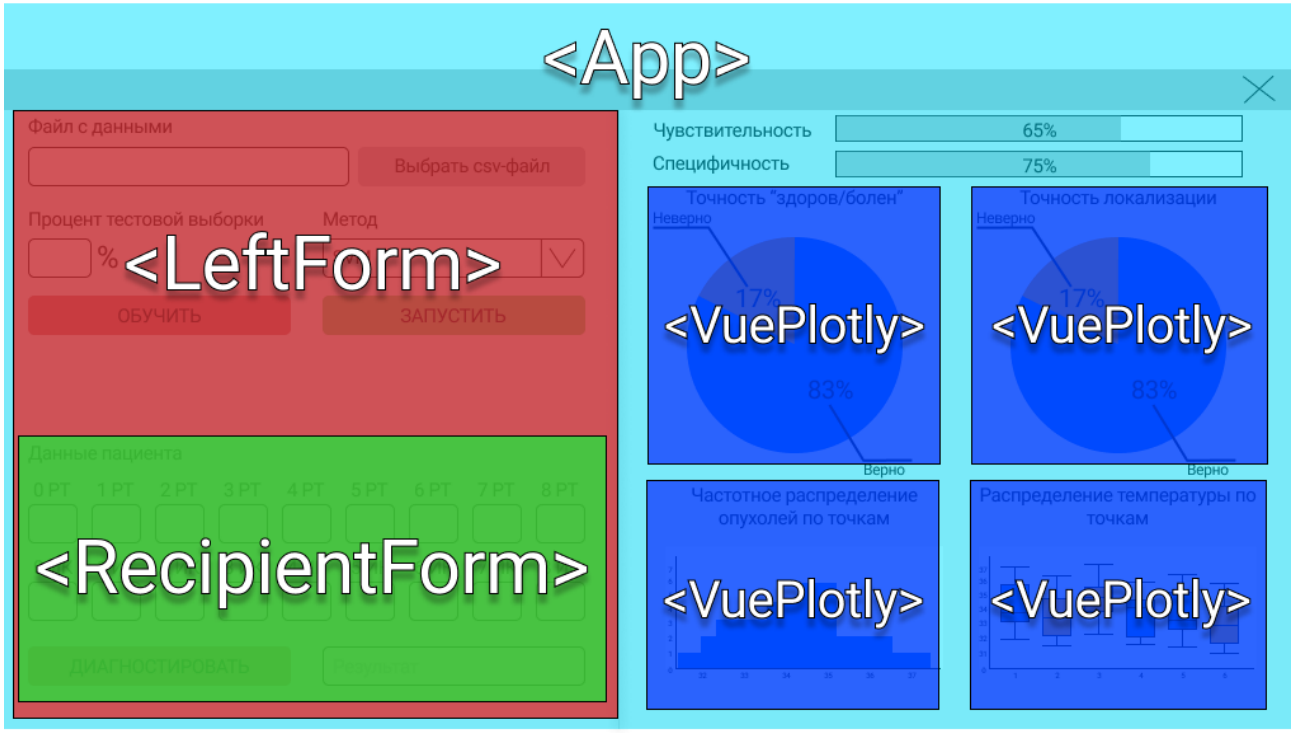

Рисунок 12 – Структура Vue-компонентов

Для построения графиков была использована библиотека Plotly. Данная библиотека позволяет строить различные графики и имеет встроенный функционал для масштабирования и сохранения графиков на компьютер. Но она не имеет встроенную поддержку фреймворка VueJS. Поэтому был разработан компонент для отображения графиков с помощью Plotly.

Фронтенд получает данные отправляя AJAX-запросы к API-методам сервера, т.е. без перезагрузки страницы. Для каждого такого запроса сервер возвращает статус ответа и данные. При любом ответе пользователь увидит либо сообщение об ошибке (рисунок 13), либо с успешным статусом(рисунок 14).

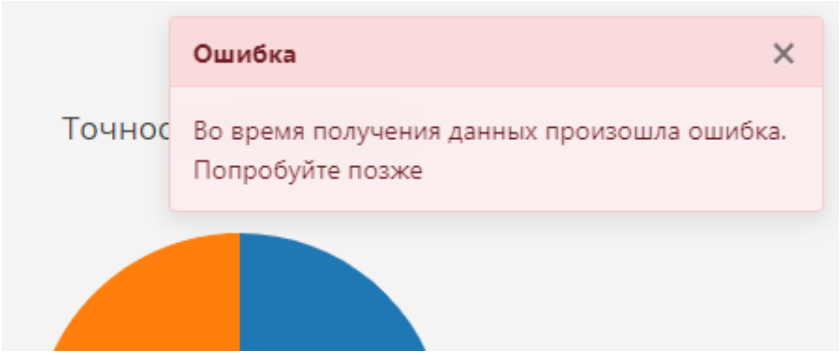

Рисунок 13 – Пример сообщения об операции с ошибкой

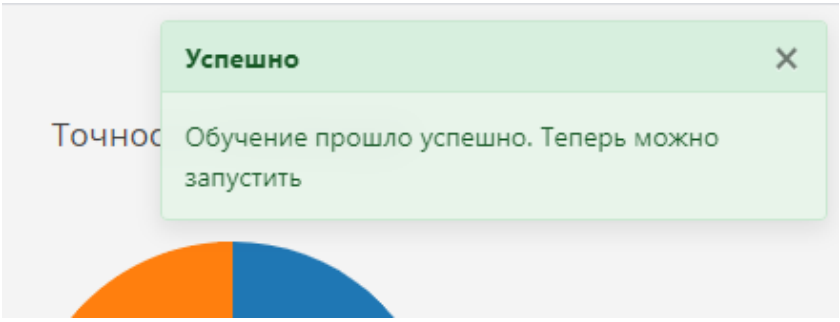

Рисунок 14 – Пример сообщения об успешной операции

Если пользователь нажмет на кнопку "Запустить но при этом модель для выбранного метода не будет обучена, то он получит уведомление как на рисунке 15).

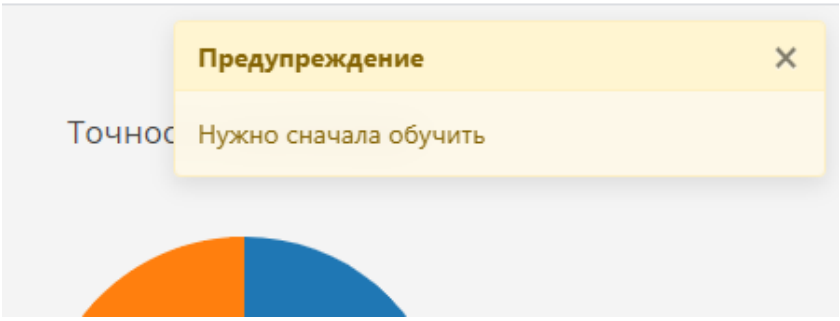

Рисунок 15 – Сообщение, если в момент запуска классификации не была найдена обученная модель

Итоговый вариант интерфейса получившейся программы представлен на рисунке 16.

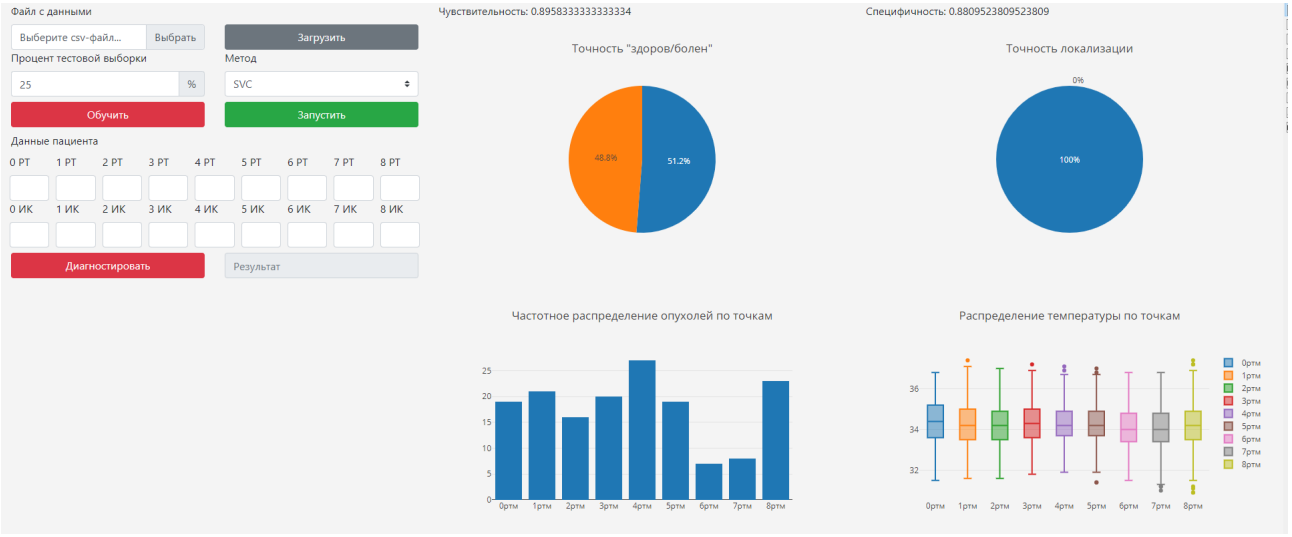

Рисунок 16 – Скриншот получившейся программы

2.4.3 Разработка Unit-тестов

Для контроля правильности работы уже существующих функций и методов приложения во время добавления нового функционала были разработаны Unit-тесты.

Unit-тесты – это набор скриптов, которые в автоматическом режиме проверяют результат работы наиболее частых вариантов вызовов функций с заранее известным результатом. Такой подход широко используется при разработке программного обеспечения. Тесты пишутся разработчиком и запускаются перед переносом нового функционала в «боевое» окружение.

В рамках данной работы были реализованы Unit-тесты для методов расчета чувствительности и специфичности. Так же был разработан набор тестов для методов API, где проверяются статусы и тело ответов при отправке запросов.

Запуск тестов производится с помощью пакета PyTest и запускается командой pytest в командной строке, либо это можно настроить в IDE, в которой происходит разработка. На рисунке 17 показан пример с результатом запуска тестов в IDE PyCharm от компании JetBrains.

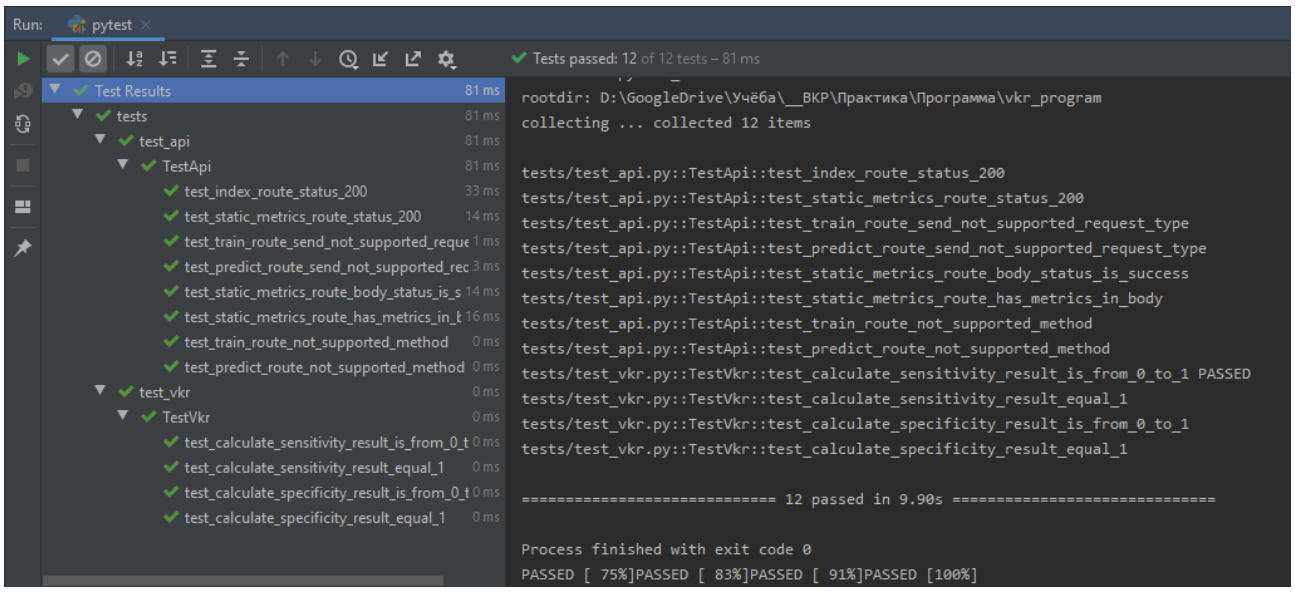

Рисунок 17 – Результат запуска Unit-тестов в IDE PyCharm

3 Применение методов машинного обучения для классификации данных компьютерного моделирования

Работа получившейся программы была протестирована на тестовом наборе данных из общей выборки(рисунок 6). Рассмотрим детальнее каждый из этапов по работе с программой.

3.1 Обучение модели и классификация тестовой выборки

Сначала нужно загрузить CSV-файл с результатами компьютерного моделирования биотканей. Для этого нужно нажать на кнопку «Выбрать» возле специального поля и выбрать файл на компьютере(рисунок 18).

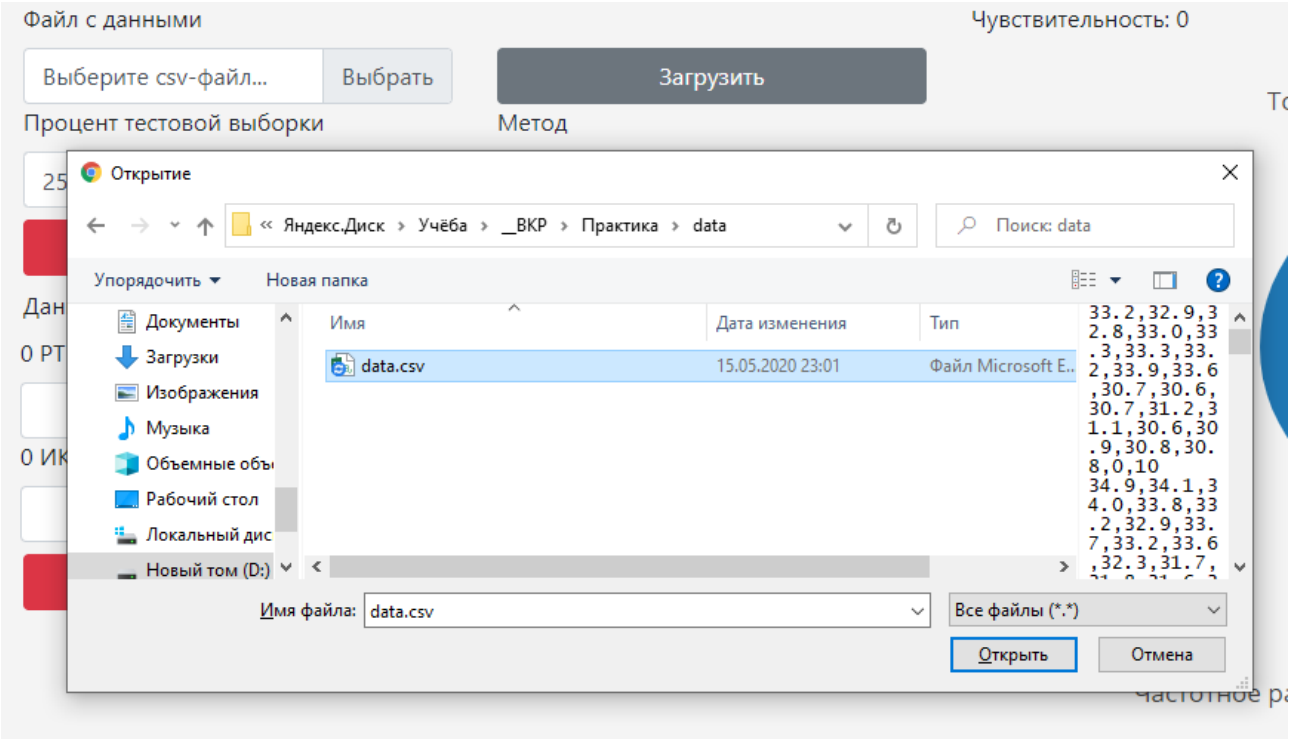

Рисунок 18 – Скриншот интерфейса программы при выборе файла с данными для загрузки

После выбора файла и нажатия на кнопку «Загрузить» файл будет загружен на сервер. Если файл загрузился успешно, то пользователь увидит

сообщение как на рисунке 19. После загрузки произойдет переинициализация всех используемых в приложении методов библиотеки Scikit-learn и будут очищены файлы с сохраненными моделями.

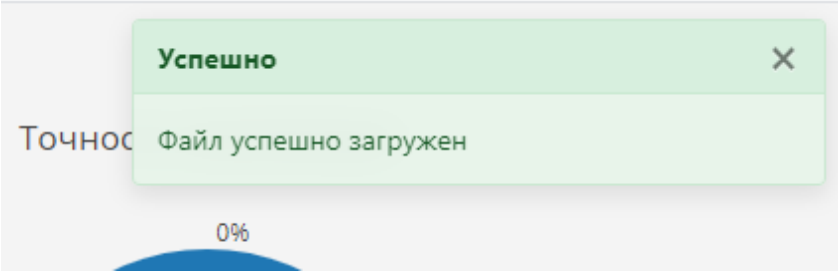

Рисунок 19 – Сообщение об успешной загрузке файла на сервер

Также в результате обновления файла были обновлены данные статистических метрик и отрисованы графики (рисунок 20). Исходя из этих данных можно сделать первые выводе об используемых при обучении данных.

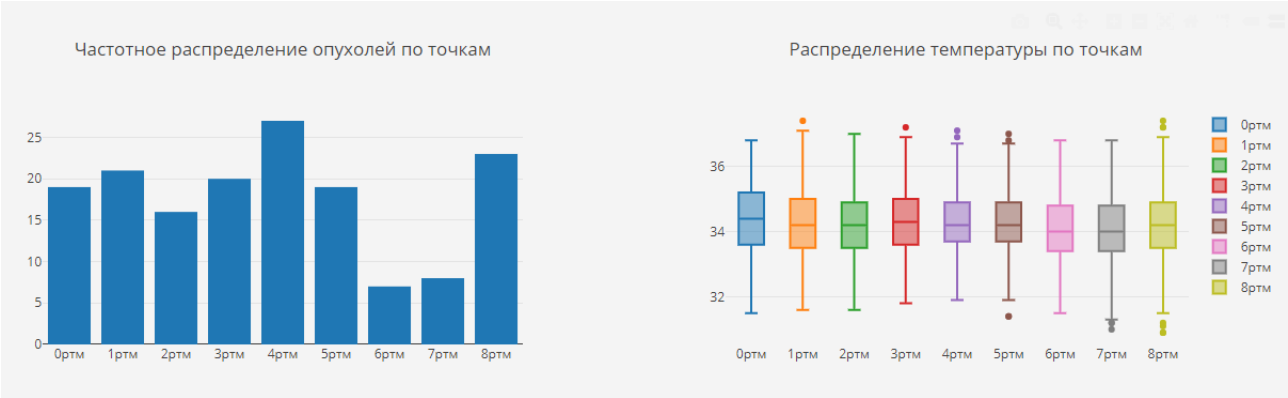

Рисунок 20 – Скриншот графиков со статистическими метриками, вычисленных на основе загруженных данных

На примере графика с частотным распределением опухолей по точкам можно заметить, что больше всего опухолей расположено в крайних точках с номерами 4 и 8, а меньше всего моделей с опухолями в соседних точках с номерами 6 и 7.

Следующим этапом после загрузки файла следует выбор метода классификации из списка, обучение модели и запуск классификации для тестовой выборки. Для всех перечисленных действий есть отдельные элементы управления (рисунок 21).

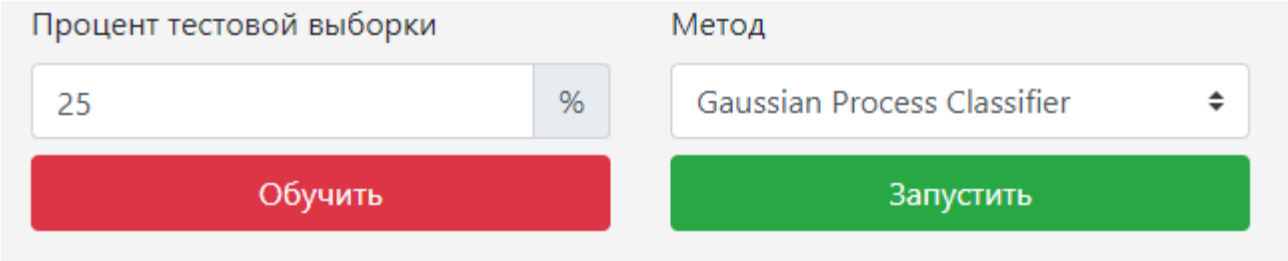

Рисунок 21 – Элементы управления для выбора размера тестовой выборки, обучения модели и классификации тестовой выборки

3.2 Определение класса «Болен»/«Здоров» и точки с опухолью

Результатом обучения и классификации является круговая диаграмма с точностью определения класса «Болен»/«Здоров». На рисунке 22 изображен пример круговой диаграммы с точностью для наивного байесовского классификатора.

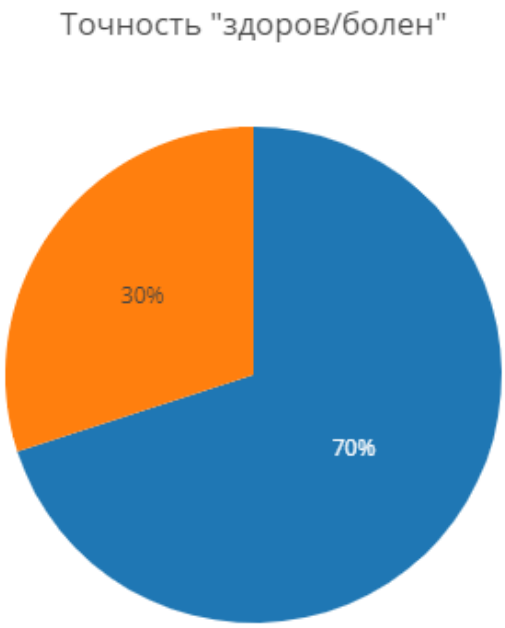

Рисунок 22 – Круговая диаграмма с точностью классификации для наивного байесовского классификатора

После тестирования нескольких методов классификации были получены следующие результаты по точности определения класса:

- SVM – 66.7%;
- k-ближайших соседей – 65.4%;
- Наивный байесовский классификатор – 70%;
- Ансамблевый метод bagging и SVM – 68.8%;
- Стохастический градиентный спуск – 56.3%.

Исходя из представленных результатов, можно сделать вывод, что для данной выборки лучше всего отработал наивный байесовский классификатор и ансамблевый метод bagging в сочетании с SVM.

Так же есть возможность определить класс (диагноз) по данным пациента. После заполнение всех нужных полей и нажатия на кнопку «Диагностировать» буду получены данные о классе и точки, к которой ближе всего расположена опухоль. На рисунке 23 приведен пример резульата дианостирования на температурных данных одной из моделей, не попавшей в

обучающую выборку. После определения точки с опухолью строится круговая диаграмма с точностью.

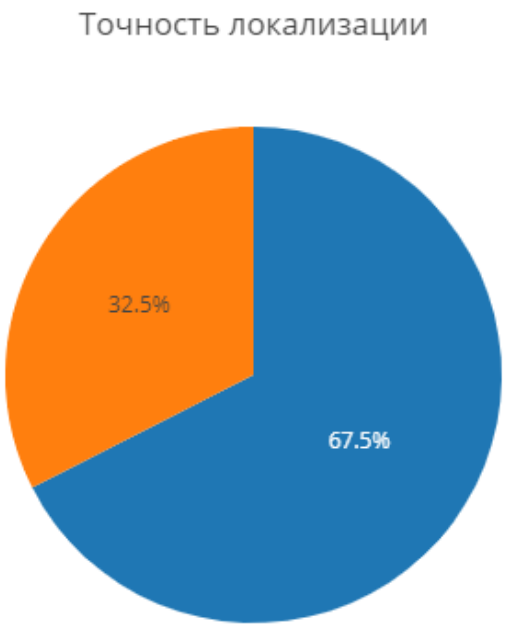

Рисунок 23 – Круговая диаграмма с точностью классификации при определении локализации опухоли

Как видно на диаграмме – получилось добиться точности определения класса равной 67.5%. Возможно этот результат получится улучшить с помощью большего объема обучающей выборки или использования другого алгоритма классификации.

Заключение

В данной работе были рассмотрены сферы деятельности и основные задачи, где используются методы машинного обучения, а также некоторые из популярных библиотек языка программирования Python для решения таких задач. Был описан принцип работы таких алгоритмов классификации как метод опорных векторов (SVM), k-ближайших соседей и наивный байесовский классификатор.

Было реализовано клиент-серверное приложение для для классификации данных компьютерного моделирования яркостной температуры. При разработке использовались такие библиотеки языка Python как Flask и Scikit-learn. При разработке клиентской части использовался фреймворк VueJs. Были разработы Unit-тесты для упрощения доработок программы в будущем. Температурные данные были разбиты на обучающую и тестовую выборки и классифицированы с помощью получившейся программы.

Исходя из результатов классификации моделей был сделан вывод, что точность классификации данных сильно зависит от используемого алгоритма. Лучше всего в проведенных экспериментах себя показал метод опорных векторов (SVM) и k-ближайших соседей.

Список литературы

1. Bardati, F. Modeling the Visibility of Breast Malignancy by a MicrowaveRadiometer / F. Bardati, S. Iudicello. – Текст : непосредственный // Biomed. Engineering. – 2008. – Vol.55 (6). – С. 214-221.

2. Cristianini, T. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods / Nello Cristianini, John Shawe-Taylor. – Текст : непосредственный // Cambridge University Press. – 2000. – 204 с.

3. Crammer, K. On the algorithmic implementation of multiclass kernel-based vector machines / Koby Crammer, Yoram Singe. – Текст : непосредственный // Journal of Machine Learning Research. – 2002. – № 2. – С. 265–292.

4. Fear, K.E. Microwave detection of breast cancer / K.E. Fear, M. Ctuchly. – Текст : непосредственный // IEEE Trans. Microwave Theory Tech. – 2000. – Vol.48 (11). – С. 1854-1863.

5. Hetal, B. An Empirical Evaluation of Data MiningClassification Algorithms / He-tal Bhavsar, Amit Ganatra. — Текст : непосредственный // International Journal of Computer Scienceand Information Security (IJCSIS). – 2016 – № 5. – С. 142–150.

6. Jian, M. A portable breast cancer detection system based on smartphone with infrared camera / Jian Ma, Pengchao Shang, Chen Lu, Safa Meraghni. — Текст : непосредственный // PROCEDIA : Vibroengineering. – 2019 – Vol. 26. – С. 57-63.

7. Kumbhar, S. Comparative Analysis of Classification Algorithms / Vijaykumar S. Kumbhar. – Текст : электронный // NCORTIT. – 2017. – 5 С. – URL: https://www.researchgate.net/publication/313440536, свободный. – Загл. с экрана.

8. Leroy, Y. Non-invasive microwave radiometry thermometry / Y. Leroy, B. Bocquet, A. Mammouni. – Текст : непосредственный // Physiol. Means. – 1998. – Vol.19. – С. 127-148.

9. Levshinskii, V. Verification and Validation of Computer Models for Diagnosing Breast Cancer Based on Machine Learning for Medical Data Analysis / V. Levshinskii, M. Polyakov, A. Losev, A. Khoperskov. – Текст : электронный // Communications in Computer and Information Science. – 2019. – Vol. 1084. – С. 447-460. – URL: https://link.springer.com/chapter/10.1007%2F978-3-030-29750-3_35, свободный. – Загл. с экрана.

10. Mirmozaffari, M. Data Mining Classification Algorithms for Heart Disease Prediction / Mirpouya Mirmozaffari, Alireza Alinezhad, Azadeh Gilanpour. – Текст : непосредственный // n-ternational Journal of Computing Communications & InstrumentationEngg (IJCCIE). – 2017. – 4, № 1 – С. 11-15.

11. Mossina, L. Naive Bayes Classification for Subset Selection / Luca Mossina, Emmanuel Rachelson. – Текст : электронный // Physiol. Means. – 1998. – Vol. 19. – С. 127-148. – URL: https://www.researchgate.net/publication/318560282, свободный. – Загл. с экрана.

12. Nisreen, I. Machine Learning Techniques for Breast Cancer Computer Aided Diagnosis Using Different Image Modalities: A Systematic Review / Nisreen I. Yassin, Shaimaa Omran, Enas M. F. El Houby, Hemat Allam. – Текст : электронный // Computer Methods and Programs in Biomedicine. – 2018. – Vol. 156. – С. 25-45. – URL: https://www.sciencedirect.com/science/article/abs/pii/S0169260717306405?via%3I свободный. – Загл. с экрана.

13. Ongeval, V. Ch. Digital mammography for screening and diagnosis of breast cancer: an overview / Ch. Van Ongeval. – Тест : непосредственный // PubMed PMID. – 2007. – Vol. 90 (3). – С. 163–166.

14. Sherwood, L. Fundamentals of Human Physiolog / L. Sherwood. – Текст : непосредственный // Belmon : Brooks/Cole – 2012. – 720 с.

15. Statnikov, A. A Gentle Introduction to Support Vector Machines in Biomedicine: Theory and methods / Alexander Statnikov, Constantin F.

Aliferis, Douglas P. Hardin. – Текст : непосредственный // World Scientific. – 2011. – 183 с.

16. Stauffer, P.R. Utility of Microwave Radiometry for Diagnostic and Therapeutic Applications of Non-Invasive Temperature Monitoring / P.R. Stauffer, D.R. Rodrigues. – Текст : непосредственный // IEEE BenMAS (Benjamin Franklin Symposium on Microwave and Antenna Sub-systems). – 2014. – С. 1-3.

17. Van Ongeval, Ch. Digital mammography for screening and diagnosis of breast cancer: an overview / Ch. Van Ongeval. – Текст : непосредственный // PubMed PMID. – 2007. – Vol. 90 (3). – С. 163–166.

18. Айвазян , С. Прикладная статистика: классификация и снижение размерности / Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. – Текст : непосредственный // Москва : Финансы и статистика, 1989. – 487 с.

19. Алгоритмы интеллектуального анализа данных. / Текст : электронный // 2015. – URL: https://tproger.ru/translations/top-10-data-mining-algorithms/, свободный. — Загл. с экрана.

20. Барсегян, А. Методы и модели анализа данных: OLAP и Data Mining / А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод. – Текст : непосредственный // Санкт-Петербург : БХВ-Петербург, 2004. – 336 с.

21. Бирюлин, Г. Теплофизический расчет в конечно-элементном пакете COMSOL/FEMLAB : методическое пособие / Г.В. Бирюлин. – Текст : непосредственный // Санкт-Петербург : СПбГУИТМО, 2006. – 75 с.

22. Большие данные и машинное обучение: новые возможности для медицины. / Текст : электронный // 2017. – URL: https://habr.com/ru/company/spbifmo/blog/340668/, свободный. — Загл. с экрана.

23. Вандер Плас, Д. Python для сложных задач. Наука о данных и машинное обучение / Дж. Вандер Плас. – Текст : непосредственный // Санкт-Петербург : Питер, 2017. – 576 с.

24. Веснин, С. Современная микроволновая радиотермометрия молочных желез / С.Г. Веснин, М.А. Каплан, Р.С. Авакян. – Текст : электронный // Маммология/Онкогинекология. – 2008. – №3 – 8 с. – URL: https://elibrary.ru/item.asp?id=11610722, свободный. – Загл. с экрана.

25. Веснин, С. Миниатюрные антенны-аппликаторы для микроволновых радиотермометров медицинского назначения / С.Г. Веснин, М.К. Седанкин. – Текст : электронный // Биомедицинская радиоэлектроника. – 2011. – №10 – С. 51-56. – URL: https://elibrary.ru/item.asp?id=17090823, свободный. – Загл. с экрана.

26. Веснин, С. Разработка серии антенн-аппликаторов для неинвазивного измерения температуры тканей организма человека при различных патологиях / С.Г. Веснин, М.К. Седанкин. – Текст : электронный // Вестник МГТУ им. Н.Э. Баумана. Сер. «Естественные науки». – 2012. – №11 – С. 43-61. – URL: https://elibrary.ru/item.asp?id=20179995, свободный. — Загл. с экрана.

27. Вьюгин, В. Математические основы теории машинного обучения и прогнозирования / Владимир Вьюгин. – Текст : электронный // МЦМНО. – 2013. – 390 с.

28. Гудфеллоу, Я. Глубокое обучение / Гудфеллоу Я., Бенджио И., Курвилль А. – Текст : электронный // Москва : ДМК Пресс. – 2017. – 652 с.

29. Данилов, С. Интеллектуальный анализ данных с использованием системы Rapid Miner / С.В. Данилов. – Текст : электронный // Казанский (Приволжский) федеральный университет. – 2014. – 43 с.

30. Дауни, А. Байесовские модели / Дауни А.Б., пер. с анг. В. А. Яроцкого – Текст : непосредственный // Москва : ДМК Пресс. – 2018. – 182 с.

31. Доусон, М. Программируем на Python / Доусон М. – Текст : непосредственный // Санкт-Петербург : Питер. – 2019. – 416 с.

32. Журавлев, Ю. «Распознавание». Математические методы. Программная система. Практические применения / Журавлев Ю. И., Рязанов В. В., Сенько О. В. – Текст : непосредственный // Москва : Фазис, 2006. – 176 с.

33. Левитин, А. Алгоритмы. Введение в разработку и анализ / Левитин А. В. – Текст : непосредственный // Москва : Вильямс. – 2006. – 576 с.

34. Лосев, А. Проблемы измерения и моделирования тепловых и радиационных полей в биотканях: анализ данных микроволновой термометрии / А.Г. Лосев, А.В. Хоперсков, А.С. Астахов, Х.М. Сулейманова. – Текст : непосредственный // Вестн. Волгогр. гос. ун-та. Сер. 1, Мат. Физ. – 2015. – No 6 – 41 с.

35. МакГрат, М. Алгоритмы. Python. Программирование для начинающих / Майк МакГрат. – Текст : непосредственный // Эксмо. – 2013. – 194 с.

36. Мюллер, А. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными / Андреас Мюллер, Сара Гвидо. – Текст : непосредственный // Вильямс. – 2017. – 480 с.

37. Николенко, С. Алгоритмы. Глубокое обучение / Николенко С., Кадурин А., Архангельская Е. – Текст : непосредственный // Санкт-Петербург : Питер. – 2018. – 480 с.

38. Паклин, Н. Бизнес-аналитика: от данных к знаниям : Учебное пособие / Паклин Н.Б., Орешков В.И. – Текст : непосредственный // Санкт-Петербург : Питер, 2013. – 2-е изд. – 704 с.

39. Поляков, М. Математическое моделирование пространственного распределения радиационного поля в биоткани: определениеяркостной температуры для диагностики / М.В. Поляков, А.В. Хоперсков. – Текст : непосредственный // Вестн. Волгогр. гос. ун-та. Сер. 1, Мат. Физ. – 2016. – No 5 – С. 73-84.

40. Потапов, М. Анализ эффективности алгоритмов интеллектуального анализа данных для решения задачи распознавания изображений со спутников / Потапов М. П. – Текст : электронный // Федеральное государственное бюджетное образовательное учреждение высшего образования "Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева": Актуальные проблемы авиации и космонавтики, 2016. – 1. – № 12 – С. 563-565.

41. Рамсундар, Б. Глубокое обучение в биологии и медицине / Рамсундар Б., Истман П., Уолтерс П., Панде В. – Текст : непосредственный // Москва : O'Reilly, 2019. – 202 с.

42. Рашка, С. Python и машинное обучение / Рашка С., пер. с англ. А. В. Логунова. – Текст : непосредственный // Москва : ДМК Пресс, 2017. – 418 с.

43. Розенблатт, Ф. Принципы нейродинамики: Перцептроны и теория механизмов мозга / Розенблатт Ф. – Текст : непосредственный // Москва : Мир, 1965. – 480 с.

44. Флах, П. Машинное обучение / Флах П. – Текст : непосредственный // Москва : ДМК Пресс, 2015. – 400 с.

45. Шлезингер, М. Десять лекций по статистическому и структурному распознаванию / Шлезингер М., Главач В. – Текст : непосредственный // Киев : Наукова думка, 2004. – 546 с.