ПРОГРАММИРОВАНИЕ, 2022, № 4, с. 3–16

ПРОГРАММНАЯ ИНЖЕНЕРИЯ, ТЕСТИРОВАНИЕ И ВЕРИФИКАЦИЯ ПРОГРАММ

УДК 004.421.6

ПРОГНОЗИРОВАНИЕ СТЕПЕНИ ПОРАЖЕНИЯ ЛЕГКИХ ПРИ COVID-19 НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

© 2022 г. Ю. А. Васильев[a,*] (ORCID: 0000-0001-9210-5544),
М. И. Петровский[a,] (ORCID: 0000-0002-1236-398X),**
И. В. Машечкин[a,*] (ORCID: 0000-0002-9837-585X),**
Л. Л. Панкратьева[b,**] (ORCID: 0000-0002-1339-4155)**

[a] МГУ им М.В. Ломоносова, факультет вычислительной математики и кибернетики, кафедра интеллектуальных информационных технологий, 119991, Москва, ГСП-1, ул. Колмогорова, д. 1, стр. 52, Россия

[b] Федеральный научно-клинический центр детской гематологии, онкологии и иммунологии им. Дмитрия Рогачева Минздрава России, 117997, Москва, ул. Саморы Машела, д. 1, Россия

**E-mail: iuliivasilev@gmail.com*

***E-mail: michael@cs.msu.su*

****E-mail: mash@cs.msu.su*

*****E-mail: liudmila.pankratyeva@gmail.com*

Поступила в редакцию 27.12.2021 г.
После доработки 16.01.2022 г.
Принята к публикации 29.01.2022 г.

Работа посвящена решению задачи прогнозирования течения заболевания у пациентов с COVID-19. На основе данных об анамнезе, осмотре, результатах клинико-лабораторного анализа и других факторов, потенциально связанных с тяжестью течения заболевания и вероятностью смерти пациентов с COVID-19, был разработан комплекс моделей, построенных с использованием методов машинного обучения и прикладного статистического анализа, для прогнозирования тяжести течения и исхода заболевания у пациентов, получающих лечение в амбулаторных и стационарных условиях.

Одним из ключевых результатов проведенной работы является создание сервиса “КТ-калькулятор”, встроенного в городскую медицинскую информационную систему – способа оценки степени изменения легочной ткани при COVID-19 в экспресс-режиме без использования компьютерной томографии, и позволяющего на основе физикальных и лабораторных признаков спрогнозировать степень поражения легких пациентов.

Построенные в рамках данного проекта модели машинного обучения дают возможность судить о степени риска легкой и тяжелой формы течения заболевания в зависимости от различных категорий факторов.

DOI: 10.31857/S0132347422040069

1. ВВЕДЕНИЕ

Всемирная организация здравоохранения 11 марта 2020 г. объявила пандемию по заболеванию COVID-19, вызываемому вирусом SARS-CoV-2. С начала пандемии Systems Science and Engineering (CSSE) at Johns Hopkins University[1] в Российской Федерации зарегистрировано более 10 млн случаев заражения COVID-19 и более 292 тыс. случаев летального исхода по состоянию на 20 декабря 2021 года.

Ежедневный рост количества зараженных приводит к увеличению нагрузки на врачей и медицинское оборудование, ухудшению качественной оценки состояния пациентов и увеличению затрат на здравоохранение в целом.

Пандемия резко ускорила внедрение цифровых сервисов в работу московского здравоохранения. Начиная с марта 2019 года московскими поликлиниками и стационарами был накоплен огромный объем данных по истории болезни более чем 2 млн человек.

Пациенту с подтвержденным COVID-19, проводится комплекс клинического обследования.

Во-первых, происходит сбор анамнеза, включающего индивидуальные характеристики пациента (хронические заболевания, пол, возраст и т.д.). Далее, врач проводит физикальное обследование (ЧДД, сатурация, температура, степень тя-

[1] https://origin-coronavirus.jhu.edu/map.html

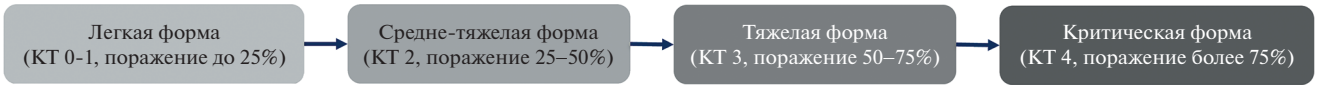

Рис. 1. Порядковое сопоставление степени поражения легких по КТ и форму COVID-19.

жести). Наконец, производится сбор анализов по направлениям на лабораторные исследования: клинический анализ крови и мочи, биохимический анализ крови.

Массовое применение оценки изменений легких получила компьютерная томография (далее – КТ). Результаты КТ могут служить предикторами необходимости госпитализации в стационар и вероятности неблагоприятного исхода в отделении интенсивной терапии. Порядковое сопоставление степени поражения легких по КТ и формы течения COVID-19 пациента представлено на рис. 1.

Однако, подход оценки изменений легких путем КТ имеет и недостатки, например, риск создания искусственных эпидемических очагов, нерациональная работа бригад скорой помощи, экономические затраты на проведение КТ. Также, возникают проблемы, связанные с радиационной безопасностью пациентов и врачей.

Актуальным является построение прогнозных моделей оценки поражения легких пациентов, основанных на собранных данных, а также выявление наиболее важных факторов, влияющих на развитие легочной пневмонии. В качестве альтернативного диагностического инструмента для оценки изменений легких в рамках данной работы предлагается исследовать и разработать методы интеллектуального анализа данных для решения задачи оценки степени изменения легочной ткани при COVID-19 в экспресс-режиме на основе физикальных и клинических признаков пациента.

Важной особенностью данной работы является использование реальных данных из нескольких источников: КТ-центров, лабораторий, поликлиник и стационаров. При сборе данных из нескольких источников возникает множество проблем: ошибки ввода и противоречия признаков, неоднородное заполнение признаков (в частности, клинические и биохимические анализы являются достаточно редкими для рядового пациента), разный объем и полнота данных. Без детального исследования и построения алгоритмов решения данных проблем не может быть достигнута большая точность прогнозных моделей.

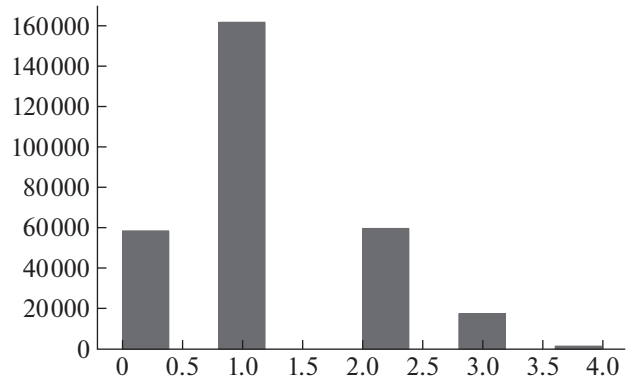

Рис. 2. Распределение результатов КТ для набора муниципальных данных.

Разработанные прогнозные модели могут быть использованы в качестве системы поддержки принятия врачебных решений, которая позволит сократить количество исследований для пациента и уменьшить облучение. Также, система снизит нагрузку на реальное оборудование – на КТ-центры и может быть применена в тех регионах, в которых доступ к компьютерному томографу ограничен, либо этого оборудования нет.

Данная работа имеет следующую структуру: в разделе 2 описываются существующие работы по применению методов машинного обучения для анализа данных COVID-19. В разделе 3 описываются особенности решаемой задачи, а также рассматривается возможность применения моделей машинного обучения. В разделе 4 описываются используемые наборы данных, метрики качества и полный алгоритм сбора и предобработки данных, включая удаление противоречий и артефактов и унификацию значений признаков. Также, в разделе описывается экспериментальное исследование оценки качества предложенных методов. В разделе 5 описывается практическая реализация сервиса “Калькулятор-КТ”, методы взаимодействия с сервисом и способы интеграции во внешнюю среду. В разделе 6 представлены основные результаты работы.

2. ОБЗОР ЛИТЕРАТУРЫ

В данном разделе рассматриваются несколько типов исследовательских работ, посвященных применению методов машинного обучения для анализа данных COVID-19, в частности, по анализу распространяемости COVID-19 [1], анализу снимков КТ [2], анализу клинических данных [3–5].

2.1. Анализ распространяемости COVID-19

На данный момент, существует несколько типов исследовательских работ, посвященных прогнозированию COVID-19.

Во-первых, существуют работы по анализу распространяемости COVID-19, прогнозированию ежедневной заболеваемости и смертности. В статье Garca-Cremades, Santi and Morales-Garca [1] решается задача анализа и прогнозирования временных рядов кумулятивной заболеваемости по COVID-19 на 14-дневный период. В работе рассматриваются модели нейронных сетей LSTM и GRU, а также статистические модели AR и ARIMA.

В качестве набора данных используются данные о мобильности в Испании, собранные с помощью инструмента Google (GMD)[2]. Он показывает набор агрегированных и анонимизированных данных, полученных от продуктов Google (в частности, Google Maps). Основными признаками набора данных являются тенденции мобильности граждан по различным категориям мест: парки, супермаркеты, общественный транспорт, рабочие места, жилье.

Лучшие результаты прогнозирования в краткосрочном периоде показал ансамблевый подход рассматриваемых методов ($0.93R^2$, $4.16RMSE$, $1.08MAE$).

2.2. Анализ снимков КТ

Во-вторых, существуют работы по анализу снимков КТ и прогнозированию степени тяжести течения COVID-19. В статье Elaziz, Mohamed Abd and Hosny [2] решается задача бинарной классификации рентгеновских снимков грудной клетки по наличию COVID-19 у пациента. Предлагаемый способ извлекает особенности из рентгеновских снимков грудной клетки с использованием ортогональных многочленов с дробными порядками FrMEMs.

Далее, используется алгоритм отбора признаков MRFODE, заключающийся в генерации набора решений и вычисления значения пригодности для каждого из признаков с использованием классификатора KNN (K ближайших соседей) на основе обучающего набора с определением наилучшего из признаков. Процесс выбора лучшего признака производится до достижения предельных условий. На основе дифференциальной эволюции (DE) генерируются бинарные вектора по выбранным признакам и обучается классификатор KNN.

В работе рассматриваются два разных набора данных. Первый набор данных, собранный по педиатрическим пациентам в возрасте от одного до пяти лет из медицинского центра Гуанчжоу, содержит изображения обычной и вирусной пневмонии. Набор содержит 216 положительных и 1675 отрицательных изображений COVID-19. Второй набор данных представляет выгрузку из базы данных Итальянского общества медицинской и интервенционной радиологии (SIRM) по пациентам с COVID-19. Набор данных состоит из 219 положительных и 1341 отрицательного изображения COVID-19.

Экспериментальное исследование качества методов проводилось по метрике Accuracy и включало сравнение предложенного метода с обученной глубокой нейронной сетью MobileNet. Предложенный алгоритм показал лучшие результаты и достиг показателей точности 0.9609 и 0.9809 для первого и второго наборов данных соответственно.

2.3. Анализ клинических данных

Наконец, существуют работы по анализу клинических данных госпитализированных пациентов и прогнозированию смертности от COVID-19. В статье Levy, Todd J. and Richardson [3] решается задача прогнозирования 7-дневной выживаемости у пациентов, госпитализированных с COVID-19. Предлагаемый метод основан на предварительном отборе наиболее значимых признаков с использованием регрессии LASSO и прогнозированием 7-дневной выживаемости на основе теоремы Байеса.

В качестве набора данных использовалась выгрузка пациентов, госпитализированных в период с 1 марта по 6 мая 2020 года, из 13 неотложных больниц Нью-Йорка. Набор содержит более 11000 пациентов со средним возрастом 65 лет и общей 7-дневной выживаемостью 89%. На основе электронной медицинской карты были извлечены 42 признака, включая демографические, лабораторные и клинические данные пациентов.

Разработанный калькулятор NOCOS (Northwell COVID-19) основан на 6 наиболее значимых признаках: азот мочевины сыворотки крови, возраст, абсолютное количество нейтрофилов, ширина распределения эритроцитов, насыщение кислородом и натрий, и на тестовой выборке достигает значения AUC 0.86.

В статье Jin, Jin and Agarwala [4] решается задача прогнозирования риска смертности от COVID-19. Предлагаемый подход основан на построении множества моделей пропорциональных рисков Кокса на подвыборках, построенных на основе местоположения и возрастной группы (рассматриваются группы 18–44, 45–74, 75+). На основе расчитанных рисков, для каждого местоположения также строится модель логистической регрессии.

[2] https://www.google.com/covid19/mobility/

В качестве набора данных использовалась выгрузка по пациентам, застрахованным в системе National Health Insurance Scheme с 7 июня 2020 года по 1 октября 2020 года. Набор данных содержит более 4.1 млн пациентов по 259 округам США, а также более 15 признаков, включая анамнез, осмотр и клинические данные. Предложенный метод достигает на тестовой выборке значения AUC 0.895.

В статье Yadaw, Arjun S. and Li [5] решается задача прогнозирования смертности госпитализированных пациентов с диагностированным COVID-19. В работе рассматриваются модели машинного обучения: Logistic Regression, Support Vector Machine, Random Forest и XGBoost.

В качестве набора данных используется выгрузка по пациентам, проходивших лечение в Системе здравоохранения Маунт-Синай в Нью-Йорке, штат Нью-Йорк, США. Набор содержит более 3800 наблюдений и более 20 признаков, включая анамнез (возраст, пол, хронические заболевания) и осмотр врача (сатурация, артериальное давление, температура).

Лучшие результаты прогнозирования смертности пациентов по метрике AUC показал метод XGBoost (0.91 AUC), основанный на идее градиентного бустинга и фокусирующийся на более сложных для прогнозирования подмножествах обучающих данных. Наибольшую значимость в прогнозных моделях получили признаки: сатурация, возраст, артериальное давление.

3. ИССЛЕДОВАНИЕ И ПОСТРОЕНИЕ РЕШЕНИЯ ЗАДАЧИ

Задача прогнозирования степени поражения легких по КТ пациентов может быть представлена в виде решения двух задач бинарной классификации:

1. Определение вероятности легкой степени поражения (КТ 0–1) – пациент не требует госпитализации и может проходить лечение на дому.

2. Определение вероятности тяжелой степени поражения (КТ 3–4) – пациент должен быть незамедлительно госпитализирован в стационар для проведения интенсивной терапии без промежуточного посещения КТ-центра или поликлиники.

3.1. Особенности решения задачи

Важной особенностью данной работы является использование реальных медицинских данных, имеющих несколько значимых проблем.

Во-первых, возникает проблема сложности сбора данных: для составления полной картины лечения пациента требуется обработать данные из нескольких источников (КТ-центр, поликлиника, стационар, лаборатории клинических анализов и тестов). Каждый источник обладает разным объемом данных и может содержать различные ошибки ввода.

При агрегации данных из различных источников, могут возникать противоречия в данных и пропуски в признаках пациентов. Для корректной работы моделей прогнозирования требуется предварительно провести удаление артефактов, противоречий и исследовать подходы для обработки пропусков.

Из-за различного объема данных в источниках также возникает проблема несбалансированности заполнения признаков пациентов. Например, признаки из КТ-центров будут заполнены у всех пациентов, так как целевым признаком является результат степени поражения по КТ. А признаки клинических или биохимических анализов могут быть не заполнены (по причине не проведения тестов или потери данных) или быть неактуальными из-за давнего срока сдачи.

Также, возникает проблема полноты данных, так как доступ к некоторым данным ограничен в силу сложности сбора. В частности, в медицинских исследованиях [6] отмечается значимость признака SpO2 (сатурация) на течение COVID-19. Однако, в доступных источниках данных сатурация встречается у малого числа пациентов. Для повышения заполненности признаков могут быть задействованы дополнительные источники данных, а также произведена аппорксимация на основе имеющихся признаков.

Наконец, существует проблема насбалансированности набора данных относительно степени поражения легких по КТ. Доминирующее большинство пациентов имеют легкую степень поражения КТ-1, а критические степени КТ-3 и КТ-4 имеются у малой доли выборки. Многоклассовая классификация с балансировкой классов приведет к использованию только части имеющихся данных. В данной работе предлагается разработка одноклассовых классификаторов высокой и легкой степени, каждый из которых построен на основе сбалансированных выборок по степени поражения по КТ.

Данный раздел содержит описание базовых моделей машинного обучения с учителем [7, 8], в частности, бинарные классификаторы: метод случайного леса [9, 10], нейронные сети [11, 14] и градиентный бустинг [15, 16].

Выбор данных моделей обуславливается наличием сложных нелинейных зависимостей в данных. Однако, так как заполненность признаков набора данных неравномерна, при использовании базовых моделей может наблюдаться нестабильность работы моделей и переобучение.

Для повышения устойчивости прогнозов базовых моделей и учета неоднородной заполненно-

сти признаков, в данном разделе также рассматриваются ансамбли базовых моделей прогнозирования.

3.2. Метод случайного леса (Random Forest)

Предложенный в статье [9], алгоритм Random Forest основан на идее построения ансамбля деревьев решений [10] и агрегации их прогнозов:

1. Строится N bootstrap выборок (с возвращением) из исходной выборки. Каждая bootstrap подвыборка в среднем исключает 37% данных, которые называются out-of-bag (OOB).

2. На каждой bootstrap выборке строится дерево решений, в каждом узле дерева выбирается P произвольных признаков для поиска лучшего разбиения. Выбирается разбиение, которое максимизирует разницу между дочерними узлами (в частности, максимизует logrank статистику).

3. Деревья выживаемости строятся до исчерпания bootstrap выборки.

Классификация объектов проводится путем голосования: каждое дерево ансамбля относит классифицируемый объект к одному из классов, и побеждает класс, за который проголосовало наибольшее число деревьев.

3.3. Нейронные сети (Neural Networks)

Предложенный в статье [11], алгоритм нейронной сети имитирует функционирование человеческой нейронной системы мозга. С помощью нейронных сетей можно сколь угодно точно аппроксимировать любую непрерывную функцию и имитировать любой непрерывный автомат.

Нейрон представляет собой единицу обработки информации в нейронной сети. В этой модели можно выделить три основных элемента. Во-первых, это набор синапсов x_j, связанных с нейронами k, характеризуемых весом $w_k j$. Во-вторых, сумматор, складывающий входные сигналы, взвешенные относительно соответствующих синапсов нейрона, и добавляющий смещение b_k. Наконец, к полученной сумме применяется функция активации $\varphi(.)$, формирующая выходной сигнал нейрона y_k.

Таким образом, в математическом представлении функционирование нейрона k описывается уравнением:

$$y_k = \varphi\left(\sum_{j=1}^{m} w_{kj}x_j + b_k\right) \qquad (3.1)$$

Поскольку модель нейрона реализует функцию от его входов, нейроны можно объединять в соответствии с правилами суперпозиции функций, получая более сложные модели, называемые персептронами [12] или искусственными нейронными сетями прямого распространения.

Многослойный персептрон [13] имеет несколько отличительных признаков: каждый нейрон имеет нелинейную функцию активации, сеть содержит один или несколько слоев скрытых нейронов.

Для обучения многослойного персептрона используется метод обратного распространения ошибки (от англ. Back propagation) – алгоритм обучения, основанный на вычислении градиента функции ошибок. В процессе обучения веса нейронов каждого слоя нейросети корректируются с учетом сигналов, поступивших с предыдущего слоя, и невязки (отклонения) каждого слоя, которая вычисляется рекурсивно в обратном направлении от последнего слоя к первому.

При одноклассовой классификации в качестве функции ошибок наиболее распространена бинарная кросс-энтропия, а в качестве функции активации – логистическая функция [17].

Также, для избежания переобучения, в архитектуру нейронной сети может быть добавлен слой регуляризации, ограничивающий размер весов. На практике, наибольшее распространение получили слои dropout, зануляющие часть весов перед входом в следующий слой. При использовании слоя dropout, нейросетевая архитектура обучается на частично заполненных данных, предотвращая формирование глобальных зависимостей от малого числа признаков. Используя полную информацию доступных признаков, повышается усточивость архитектуры нейронной сети на реальных данных.

3.4. Градиентный бустинг (Gradient Boosting Machines)

Альтернативным подходом построения ансамбля деревьев решений является Gradient Boosting, представленный в статье Friedman, Jerome H. (2001) [15].

В отличии от Random Forest, алгоритм Gradient Boosting основывается не на независимом построении деревьев решений и усреднении их прогнозов, а на итеративном алгоритме обучения очередного дерева решений на ошибках предыдущего. Агрегация прогнозов деревьев основана на весовых коэффициентах, рассчитываемых при добавлении нового дерева решений в ансамбль.

Целью алгоритма является минимизация loss-функции L, на основе которой производится расчет ошибки ансамбля. По умолчанию, в качестве loss-функции используется логарифмическая функция потерь (log loss) [17]. Пусть $\{(x_i, y_i)\}_{i=1}^{n}$ – обучающий набор, L – функция потерь, M – раз-

мер ансамбля. Алгоритм представляет собой следующую последовательность шагов:

1. Модель инициализируется константным значением α:

$$F_0(x) = \text{argmin}_\alpha \sum_{i=1}^{n} L(y_i, \alpha)$$

2. Вычисляются псевдоостатки для всех наблюдений в обучающей выборке $i = 1, \ldots, n$:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$$

3. На обучающей выборке $\{(x_i, r_{im})\}_{i=1}^{n}$ строится дерево решений $h_m(x)$.

4. Вычисляется вес v_m $(0 < v_m < 1)$ дерева решений с помощью решения следующей оптимизационной задачи:

$$v_m = \arg\min_v \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + v \cdot h_m(x_i)).$$

5. В ансамбль добавляется обученное дерево с весом:

$$F_m(x) = F_{m-1}(x) + v_m \times h_m(x).$$

6. Если размер ансамбля m не равен M, то переходим на шаг 2.

7. Итоговой моделью является F_M.

Классификация объектов проводится путем композиции откликов моделей ансамбля: каждое дерево решений $h_m(x)$ возвращает вещественную "степень" принадлежности объекта к некоторому классу, а результирующий ответ F_M получается применением порогового правила к композиции.

3.5. Ансамбль случайного леса RF и нейронных сетей NN

Для раздельной обработки признаков с разной заполненностью, предлагается использовать агрегацию базовых моделей прогнозирования. Предлагается использовать ансамбль двух моделей прогнозирования: первая модель обучается на признаках с большим количеством пропусков, а вторая модель обучается на хорошо заполненных признаках и отклике первой модели.

Наибольшее количество пропусков выявлено у группы признаков лабораторных анализов. В данную группу входят все показатели общего анализа крови и биохимических анализов. На основе лабораторных признаков предлагается обучить нейронную сеть. Отклик модели добавляется к группе оставшихся признаков, на основе которой производится обучение случайного леса.

Вероятностный прогноз случайного леса используется при оценке качества модели. Выбор нейронной сети для обучения на малозаполненных признаках обуславливается лучшей аппроксимацией на большом признаковом пространстве числовых признаков (дополненного бинарными признаками наличия пропуска), которые могут коррелировать друг с другом. В случае использования древовидных ансамблей на признаках с пропусками, часть признаков могут быть вовсе не использованы или иметь малую значимость.

3.6. Бустинг ансамбль деревьев решений LGBM

LGBM (Light Gradient Boosting Machine) [16] – это одна из наиболее эффективных реализаций процедуры градиентного бустинга. Поскольку данный метод является ансамблем деревьев решений, LGBM позволяет оценивать важность признаков из обученной модели. Как правило, важность обеспечивает оценку, которая указывает, насколько полезным был каждый признак при построении деревьев решений в модели. Чем больше атрибут используется для принятия ключевых решений, тем выше его относительная важность. Важность рассчитывается для отдельного дерева решений, затем значения характеристик усредняются по всем деревьям решений в модели.

Особенность LGBM заключается в обучении только на тех данных, которые приводят к большему градиенту, что способствует ускорению работы алгоритма и уменьшению его вычислительной сложности.

Также, LGBM позволяет обрабатывать пропуски во входных признаках. Для числовых признаков пропущенные значения соотносятся той ветви разбиения, которая в наибольшей степени уменьшает функцию потерь. Для категориальных признаков пропущенные значения используются в качестве отдельной категории. В таком случае, нет необходимости использовать ансамбль нескольких видов моделей машинного обучения, обрабатывающий признаки с разной заполненностью.

3.7. Ансамбль регуляризованных нейросетей

В чистом виде метод обратного распространения ошибки работает плохо [11, 18]. Возникают проблемы медленной сходимости или расходимости, застревания в локальных минимумах функционала. Для стабилизации процесса обучения была добавлена инициализация и регуляризация слоев нейросети. Так как в качестве функции активации нейросети используется сигмоида, отклик нейросети может быть интерпретирован как вероятность.

Также, при различных исходных инициализациях *random_state*, различаются как результаты

Таблица 1. Структура набора муниципальных данных

Источник	Объем данных	Признаки пациента	Результат анализа	Доп. признаки
Данные КТ-центров	303’628	степень тяжести, ЧДД, температура тела, наличие и тип кашля	время проведения КТ, степень поражения КТ, результат КТ, наличие пневмонии	основной диагноз, сопутствующие диагнозы
Данные амбулаторных тестов	43’348’273	–	название теста, время проведения, время сбора, результат теста, референсные значения	название исследования, единицы измерения
Данные тестов ПЦР и ИФА	4’466’407	дата рождения, пол	тип теста, название и код диагноза, время проведения теста, даты получения и выдачи результатов теста, результат теста	наименование пункта сбора тестов, наименование филиала, наименование лаборатории
Данные по сатурации	661’353	–	время проведения, значение сатурации	–
Данные по госпитализированным пациентам	880’352	дата рождения, пол, группа риска	тяжесть заболевания, ОРИТ, ИВЛ, ЭКМО	дата госпитализации, причина поступления
Данные по умершим пациентам	48’415	дата рождения, пол	дата смерти, причина смерти, код смерти	связь смерти с COVID-19

сходимости нейросети, так и порог бинаризации отклика. Для повышения устойчивости прогноза частных нейросетей, был предложен подход формирования ансамбля нейросетей. Для входного наблюдения рассчитывается множество откликов по всем обученным нейросетям, отклики объединяются суммированием и нормируются. Полученный отклик принимается за отклик ансамбля.

Также, для контроля обучения нейронных сетей использовался обработчик прекращения обучения при ухудшении метрики качества. Максимальное количество эпох без возможного улучшения параметра – 10 эпох. Также, использовался обработчик уменьшения скорости обучения (learning rate) при ухудшении метрики качества.

3.8. Калибровка результатов и выбор порогов

При выполнении классификации часто требуется не только предсказать метку класса, но и получить вероятность соответствующей метки. Эта вероятность определяет уверенность в предсказании.

Распределение вероятностей может быть скорректировано, чтобы лучше соответствовать ожидаемому распределению, наблюдаемому в данных. Такая корректировка называется калибровкой моделей [19, 20] и используется для сведения откликов прогнозных моделей на вероятностную шкалу. Это важно как при интерпретации прогнозов, так и для принятия решений о внедрении моделей и анализа их работы.

Пусть y_i – эталонная вероятность наблюдения x_i, $p(x_i)$ – отклик модели. Задача методов калибровки – построить скорректированный отклик $\hat{p}(x)$.

В данной работе в качестве метода калибровки рассматривается калибровка Платта [21], которая работает путем подгонки модели логистической регрессии к оценкам классификатора:

$$\hat{p}(x) = \frac{1}{1 + \exp(a \cdot p(x) + b)} \qquad (3.2)$$

Параметры a, b определяются методом максимального правдоподобия на отложенной выборке. Калибровка Платта наиболее эффективна, когда искажение в предсказанных вероятностях имеет сигмовидную форму.

Также, возникает проблема соотнесения наблюдений к классам на основе спрогнозированной вероятности. В действительности, не все построенные модели имеют порог 0.5 для выходных вероятностей, а также, веса ошибок при “занижении” или “завышении” результата могут отличаться. В данной задаче ошибки “занижения” прогноза критичнее ошибок “завышения” и не-

обходимо максимизировать метрики Recall (полнота) или Sensitivity (чувствительность).

Экспертами может быть установлено минимальное допустимое значение Recall, на основе которого выбираются пороги, максимизирующие Precision (точность). Данный подход позволяет использовать множество частных порогов в различных организациях использования модели, однако не имеет унифицируемого подхода расчета порогов.

4. ЭКСПЕРИМЕНТЫ

4.1. Описание метрик качества

Для оценки качества моделей прогнозирования в данной работе рассматривается метрика ROC-AUC [22] – площадь под ROC-кривой.

ROC-кривая – график, отображающий соотношение между долей объектов от общего количества носителей признака, верно классифицированных как несущие признак, и долей объектов от общего количества объектов, не несущих признака, ошибочно классифицированных как несущие признак, при варьировании порога решающего правила (ошибок I рода).

Площадь под ROC-кривой AUC (англ. Area Under Curve) принимает значение от 0 до 1 и интерпретируема как вероятность того, что классификатор присвоит больший вес случайно выбранному положительному наблюдению, чем случайно выбранному отрицательному наблюдению.

4.2. Описание наборов данных

В работе рассматриваются два набора данных.

Первый набор данных представляет собой выгрузку муниципальных медицинских данных по заболеваемости клинических и амбулаторных COVID-19 пациентов города Москва с 03.2020 по 02.2021. Выгрузка по пациентам состоит из шести источников данных, структура выгрузки и содержание источников представлено в табл. 1. Все источники данных имеют общий идентификатор пациента. Выгрузка является достаточно полной, на ее основе формируется набор данных для обучения и тестирования прогнозных моделей.

Второй набор данных госпитализированных пациентов городской больницы города Москва представляет собой выгрузку из 95 наблюдений, содержащую: идентификатор пациента, признаки анамнеза (пол, возраст, хронические заболевания), признаки осмотра врача (степень тяжести, сатурация, ЧДД, наличие одышки, слабость, заложенность, температура тела, наличие и тип кашля), результаты лабораторных тестов (PCR, WBC, PDV, MON, GRA, LYM, PLT, HGB, RBC, MPV, HCT, RDW).

В силу малого размера и явного дисбаланса в сторону степени поражения КТ 1, набор данных может использоваться только в виде тестовой выборки для оценки качества моделей.

4.3. Подготовка данных

Перед построением прогнозных моделей необходима обработка исходных медицинских данных по пациентам с диагностированным COVID-19, с целью очистки некорректных данных и приведения значений признаков к общим шкалам, а также нахождения и обработки артефактов, выбросов и противоречивых данных.

В данной работе предлагается следующий алгоритм формирования набора данных на основе множества источников.

Для каждого наблюдения из КТ-центра производится поиск ближайших (окно – неделя) тестов пациентов. Каждому пациенту сопоставляется список тестов на основе уникального идетификатора. Тесты фильтруются по рассматриваемому множеству тестов и сортируются по близости к дате проведения КТ. Выбираются ближайшие тесты с тремя признаками: значение теста, референсные значения, дата взятия теста.

Рассматриваются множества тестов общего клинического анализа крови и наиболее заполненные тесты биохимического анализа: АЛТ, альбумин, АСТ, билирубин общий, билирубин прямой, калий общий, креатинин, лактатдегидрогеназ, мочевина, натрий общий, белок общий, хлор, щелочная фосфатаза, относительное количество нормобластов.

Производится фильтрация исходного набора данных по рассматриваемым тестам. Таким образом, производится поиск теста (а не анализа) и возможно получение результатов теста от нескольких источников исследований. Также, тесты от разных исследований объединяются в один признак теста и унифицируется размерность.

Выделяются наиболее заполненные признаки среди близких тестов. Например, для тромбоцитов существует множество признаков после объединения тестов от разных исследований: общий объем тромбоцитов в крови (тромбокрит, PCT), количество тромбоцитов, средний объем тромбоцитов в крови, ширина распределения тромбоцитов по объему. Данные признаки коррелируют между собой, поэтому в ходе формирования признакового пространства выбирается наиболее заполненный признак.

Сформированный набор данных дополняется тестами: ПЦР, ИФА, сатурация (для каждого пациента выделяются ближайшие тесты с окном – неделя). Также набор дополняется признаками хронических болезней. Каждому пациенту сопоставляется список диагнозов в формате кодов

Таблица 2. Соответствие результата КТ и степени поражения КТ

	Нулевая	Легкая	Средн.	Тяжел.	Крит.
0	**536**	1423	409	78	0
1	510	**157298**	2578	126	7
2	48	3208	**56516**	463	8
3	12	402	1285	**15814**	9
4	3	30	37	169	**1196**

МКБ-10, из которого выделяются хронические диагнозы: ишемическая болезнь сердца (I11, I20, I24, I25, I51), артериальная гипертензия (I10, O10-13, G97, I27, K76, P29, I15), сахарный диабет (G63, E10-14, H36, M14, G59, E23, N08, O24), ожирение (E66).

4.3.1. Унификация значений признаков

В данном разделе рассматривается задача очистки и приведения значений показателей к общим шкалам и словарям, включая учет референсных значений.

В ходе обработки признаков набора муниципальных данных, были построены несколько правил унификации значений признаков. Значения, выходящие за допустимые границы признака, удаляются. Например, в данных из КТ-центров были найдены значения температуры: 3.0, 3.6, 3.8, а также значения ЧДД: 0, 1 и более 150.

Для унификации единиц измерения значений признака предложен алгоритм приведения единиц измерений к наиболее частой (целевой). Изначально, выделяются префиксы и постфиксы целевой единицы измерения. Далее, для остальных единиц измерения вычисляется расстояние до целевой. Обрабатываются префиксы и постфиксы “м, мк, н, мл, к, М”, а также степенные показатели. Неунифицируемые единицы измерения удаляются.

Для унификации категориальных признаков были составлены словари всех возможных категорий.

Для унификации непрерывных признаков прозводится удаление незначащих символов с преобразованием к вещественному виду (в случае множественного значения производится разбиение по разделителям и выбирается первое значение).

Для унификации референсных значений признака были выделены 2 основных типа референсных значений: интервал $x - -y$ и полуинтервалы $< x$, $>y$. Все референсные значения приводятся к данным типам или считаются некорректными.

Для унификации результатов теста ИФА созданы 4 признака: 2 вещественных значения IGG, IGM и 2 бинарных показателя обнаружения (IGG > 10, IGM > 2.0).

4.3.2. Обработка аномальных значений

В данном разделе рассматривается задача поиска, удаления или исправления артефактов, выбросов и противоречивых данных.

В ходе обработки признаков набора муниципальных данных, было найдено несколько типов противоречивых данных. Выявление противоречий проводилось как на основе анализа признаков, так и с помощью экспертной оценки.

Во-первых, наблюдается несогласованность полей “время определения температуры при проведении КТ” и “время проведения КТ” – между датами может пройти несколько дней. В 299 тысячах наблюдений данной проблемы не наблюдается, однако более 3800 наблюдений имеют ненулевое расстояние между датами, в частности, 1 день имеют 2142 наблюдения, более двух дней имеют 1694 наблюдения, более семи дней имеют 763 наблюдения. Максимальная разница во времени составила 176 дней. В данной работе считается некорректным расстояние более 7 дней, такие наблюдения удаляются из набора данных.

Во-вторых, наблюдается несогласованность полей “степень поражения КТ” и “результат КТ”. Согласно описанию исходного набора данных, корректно следующее сопоставление между категорией КТ и степенью поражения КТ: КТ-0 – нулевая, КТ-1 – легкая, КТ-2 – средне-тяжелая, КТ-3 – тяжелая, КТ-4 – критическая. Однако, в наборе данных такому сопоставлению отвечают только 167 тысяч наблюдений. Полное соответствие категории КТ и степени поражения КТ представлено в табл. 2. Также, в 61 тысяче наблюдений значение степени поражения не указано. В данной работе будем считать, что пациент имеет степень поражения КТ-N, если его “результат КТ” КТ-N, а “степень поражения КТ” не ниже КТ-N или не указана.

В-третьих, с помощью экспертной оценки найдено противоречие, основанное на неприводимых единицах измерения. В признаках абсолютного количества эозинофилов, базофилов, моноцитов, гранулоцитов, лимфоцитов, нейтрофилов существует 2 категории значений, в процентном и количественном содержании. Данные категории невозможно преобразовать друг в друга общим подходом преобразования единиц измерения. Для корректного преобразования требуется умножить значения процентного содержания на количество лейкоцитов (WBC). При их отсутствии, значение принимается пустым.

Наконец, с помощью экспертной оценки найдено противоречие, основанное на отсутствии единиц измерения у признаков RDW, PDW, D-димер. Для устранения данной проблемы было принято решение анализировать референсные значения признаков.

Для признака RDW (и PDW) определено следующее правило: если правая референсная граница меньше 20 или написание содержит запятую, то значение представлено в виде **RDW-CV** (и **PDW-CV**) и измеряется в %, иначе представлено в виде **RDW-SD** (и **PDW-SD**) и измеряется в фемтолитрах. Сопоставление единиц измерения производится по формуле $RDW - CV = \frac{RDW - SD}{MCV} \times 100$, где MCV – средний объем эритроцитов (или по формуле $PDW - CV = \frac{PDW - SD}{MPV} \times 100$, где MPV – средний объем тромбоцитов).

Для признака D-димера, определено следующее правило: если правая референсная граница меньше 1, то значение измеряется в мкг/мл, иначе измеряется в нг/мл. Сопоставление единиц измерения производится по правилу: 1 мкг/мл = = 1000 нг/мл.

4.3.3. Обогащение набора данных

В данном разделе рассматривается задача расширения признакового пространства за счет включения информации о датах проведения анализов и осмотра.

Как было описано ранее, при формировании набора данных наблюдения с проведенной КТ обогащаются клиническими анализами, тестами ПЦР и ИФА, а также сатурацией, взятыми за недельный период от даты проведения КТ.

На практике данные анализов не всегда являются актуальными. Для контроля актуальности данных были добавлены 2 признака: дата общего анализа крови и дата биохимического анализа. На их основе вычисляется количество дней между анализами и датой КТ.

На основе данных признаков можно проводить фильтрацию актуальности входных данных, а при количестве дней, меньшем 7, использовать признак в прогнозной модели.

Также, для каждого признака N набора данных формируется признак $none\_N$, отображающий наличие в оригинальном признаке пропусков (1, если признак имел пропуск, 0 иначе).

4.3.4. Доопределение сатурации в данных

В данном разделе рассматривается задача взаимообратного учета показателя сатурации в оценке степени тяжести осмотра с пересчетом по баллам шкалы NEWS2.

При прогнозировании ожидаемой степени поражения КТ одним из важнейших факторов является показатель сатурации (насыщение крови кислородом в процентном соотношении). Однако, в сформированном наборе данных признак сатурации заполнен лишь на 103 тысячах наблюдений (из 299 тысяч).

Для увеличения заполненности по сатурации предлагается использовать международную шкалу NEWS2 [23] (National Early Warning Score), предложенную в 2020 году для оценки тяжести течения COVID-19 Королевским колледжем врачей (Royal College of Physicians).

На основе шкалы NEWS2 могут быть сформированы признаки b_spo2, bw_spo2. Признак $b\_spo2 = \frac{97 - SpO2}{2}$ определяет баллы по заполненной сатурации, а признак bw_spo2 определяет ожидаемые баллы за сатурацию на основе весовой схемы по заполненным признакам ЧДД, температуры тела и степени тяжести осмотра пациента.

Таким образом, для каждого наблюдения ставится балл за сатурацию при ее наличии или ожидаемые баллы при отсутствии.

4.3.5. Формирование выборок

Таким образом, с учетом описанной подготовки данных, по источникам муниципальных данных был сформирован набор данных с 299’792 наблюдениями. Распределение целевого поля “результат КТ” представлено на рис. 1.

Общий анализ крови заполнен на 176’354 наблюдениях (будем понимать наличие значения признака “Гематокрит”). Биохимический анализ крови заполнен на 176’866 наблюдениях (будем понимать наличие значения “С-реактивный белок”). Общий анализ крови и биохимический анализ одновременно заполнен на 151’532 наблюдениях.

После формирования исходных выборок для классификаторов КТ 01-234 и КТ 012-34, производится заполнение пропусков по медианному значению признака на исходных выборках. Также, производится балансировка классов: меньший класс размера N входит в выборку полностью, а из большего класса выбираются произвольным образом N наблюдений.

На основе полученного набора данных формируются выборки для обучения, валидации и тестирования моделей.

Для классификатора КТ 01-234 общее количество наблюдений составило 67–648 с пропорцией наблюдений в обучающей, валидационной и тестовой выборках: 54–118/6’765/6’765.

Для классификатора КТ 012-34 общее количество наблюдений составило 12’576 с пропорцией наблюдений в обучающей, валидационной и тестовой выборках: 10’062/1’257/1’257.

Также, в работе рассматривается 2 подхода формирования тестовых, валидационных и обучающих выборок. В случае рандомизированного разбиения, в обучающую выборку случайным образом определяются 80% наблюдений оригинальной выборки, а в тестовой и валидационной выборках оставшиеся наблюдения распределены в равных долях.

Таблица 3. Таблица результатов классификации КТ 01-234 и КТ 012-34

	Рандомизированная выборка		Выборка по времени	
	КТ 01-234	КТ 012-34	КТ 01-234	КТ 012-34
NN на лабораторных признаках	0.7954	0.8401	0.7899	0.7984
NN на анализах крови + RF + Калибровка Платта	0.8842	0.9227	0.8317	0.8737
Ансамбль NN	0.9153	0.9327	0.8729	0.8904
LGBM без пропусков + калибровка Платта	**0.9173**	**0.9455**	**0.8743**	**0.8988**
LGBM с пропусками + калибровка Платта	0.9170	0.9453	0.8742	0.8987

В случае разбиения по времени, в тестовую выборку определяется первые 10% наблюдений за весь период с максимальным временем проведения КТ, а в валидационную выборку последующие 10% наблюдений. Остальные наблюдения относятся к обучающей выборке.

4.4. Постановка экспериментов

Первоначально проводится предобработка набора данных, формирование признаковых пространств и целевых переменных для задач классификации КТ 01-234 и КТ 012-34.

Далее, проводится обработка признаков, заполнение пропусков на основе медианы по выборке, балансировка классов. Порождаются тестовые, валидационные и обучающие выборки на основе сформированных наборов данных.

На обучающей выборке с подбором параметров на валидационной выборке производится обучение бинарных классификаторов: нейронная сеть (NN), нейронная сеть со случайным лесом (NN+RF), ансамбль NN, LGBM с пропусками, LGBM без пропусков (пропуски заполняются медианным значением по исходной выборке классификатора).

Для построенных моделей рассчитывается прогноз на тестовой выборке, после чего проводится оценка качества прогнозирования по метрике ROC AUC.

4.5. Таблицы результатов

В табл. 3 представлена информация о полученных значениях метрики ROC AUC на рандомизированной и временной тестовой выборке для задач классификации КТ 01-234 и КТ 012-34.

Исходя из таблицы результатов, лучшие значения метрики ROC AUC на двух задачах классификации по рандомизированной и временной выборке показала модель LGBM без пропусков с калибровкой Платта.

Также, проводится тестирование модели на наборе данных городской больницы. На задаче классификации КТ 01-234 получен ROC AUC 0.8805, а на задаче классификации КТ 012-34 ROC AUC 0.9311. ROC кривые классификации изображены на рис. 3.

5. ПРАКТИЧЕСКАЯ РЕАЛИЗАЦИЯ

На основе результатов теоретических и экспериментальных исследований был разработан вебсервис “Калькулятор КТ”, предоставляющий пользователю возможность экспресс-оценки изменений легочной ткани при COVID-19 без применения компьютерной томографии органов

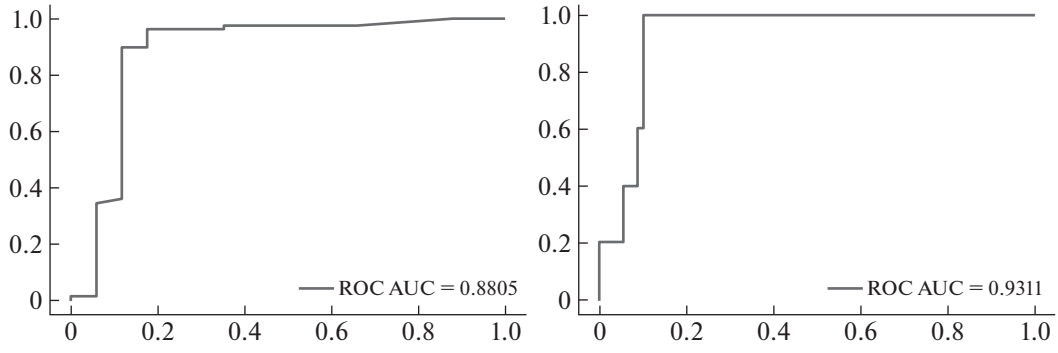

Рис. 3. ROC-кривые классификации КТ 01-234 и КТ 012-34 тестового набора данных городской больницы.

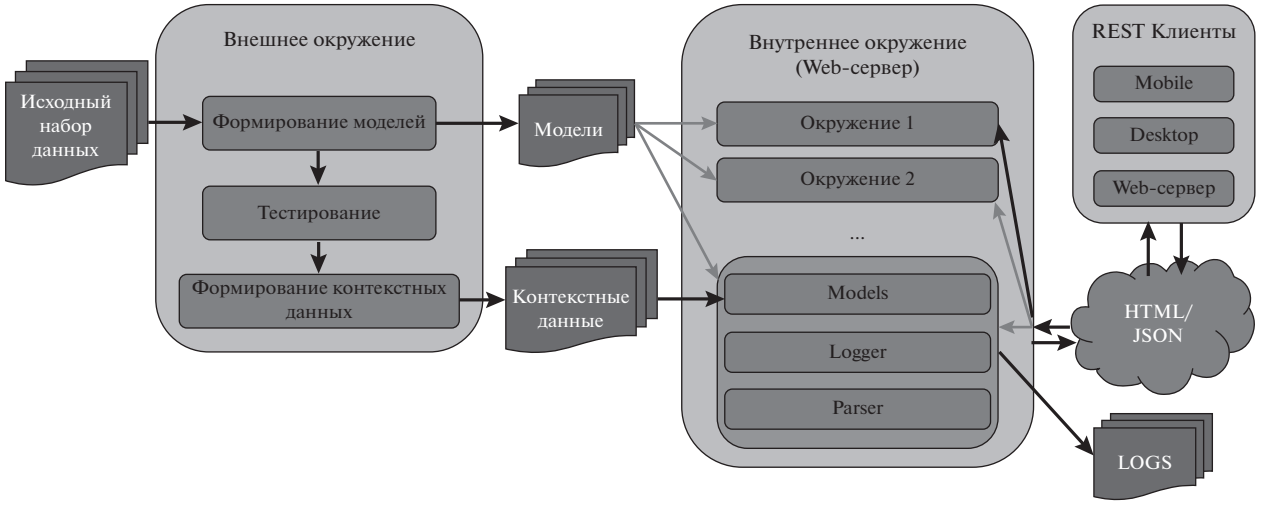

Рис. 4. Архитектура сервиса Калькулятор КТ.

грудной клетки на основе физикальных и лабораторных анализов пациента. Сервис позволяет получить прогноз вероятности легкой и тяжелой степеней поражения легких.

5.1. Архитектура реализации

Программная реализация сервиса "Калькулятор КТ" основана на двух независимых частях: внешнем и внутреннем окружении (полная архитектура представлена на рис. 4).

Внешнее окружение предназначено для детальной настройки прогнозных моделей, принимает на вход исходный набор данных, проводит предобработку признаков и формирует обучающие и тестовые выборки. На основе сформированных выборок производится обучение и выбор лучшей прогнозной модели на основе метрики ROC AUC. Тестовая выборка используется для формирования отчетов о прогнозной модели и расчета пороговых значений. Наконец, выводом внешнего окружения является файл лучшей прогнозной модели, а также контекстная информация (медианные значения по набору данных и коэффициенты стандартизации признаков).

Внутреннее окружение предназначено для запуска Web-сервиса и принимает на вход файлы, порожденные внешним окружением, проводит загрузку и инициализацию моделей. Каждый используемый вид прогнозной модели порождает окружение, состоящее из модели, функции-обработчика признаков наблюдений и логгера. В запросе пользователя возможно указание конкретного окружения прогноза.

Во внутреннем окружении также находится WSGI-сервер [24] на основе Waitress, который руководит работой Web-приложения. Web-приложение обрабатывает поступающие запросы и выполняет предсказания на основе ранее обученной модели.

При внедрении сервиса во внешнюю среду, необходимо передать только программный код внутреннего окружения, а также файлы, порожденные внешним окружением. Данный подход требует малое количество оперативной памяти, так как исходный набор данных обрабатывается перед запуском сервиса, а также не требует передачи набора данных во внешнюю среду, следуя требованиям безопасности.

После применения модели прогнозирования пользователю сервиса возвращается уникальный идентификатор запроса. При получении ошибочного результата пользователь может сообщить разработчику идентификатор запроса и разработчиком будет проведен анализ логированных данных о совершенном запросе (в частности, проверка на противоречивость). Если обращение корректно, событие размечается и вносится в тестовый набор.

Также, логируемые события используются для формирования статистики работы сервиса "Калькулятор КТ". Статистика была использована для исследования входных данных и освещена в статье на Официальном сайте Мэра Москвы[3].

5.2. Методы взаимодействия

Для взаимодействия с сервисом "Калькулятор КТ" используется архитектурный стиль REST [25], функционирующий поверх протокола HTTP. Для каждой операции сопоставляется свой собственный HTTP метод: GET – получение данных;

[3] Российские врачи 10 тысяч раз воспользовались КТ-калькулятором для диагностики COVID-19: https://www.mos.ru/news/item/86015073/

POST – создание новых данных; PUT – обновление, модификация данных; DELETE – удаление данных.

В качестве пакета данных отправляется JSON [26] массив на указанный адрес сервиса. Со стороны сервиса “Калькулятор КТ” срабатывает функция-обработчик, а в зависимости от отправленных данных и текущего запроса возвращается прогноз в определенном формате.

5.3. Интеграция сервиса во внешнюю среду с помощью docker

Практическая реализация позволяет пользователю сервиса предсказать степень поражения легких на основе протоколов течения болезни. Для получения прогноза достаточно использовать веб-форму или инструменты REST API.

Однако, в таком случае, запросы отправляются в единый сервис и логи запросов консолидируются в единое хранилище. Для поддержки нескольких независимых сервисов с локальными хранилищами, в данной работе используется программное обеспечение Docker [27, 28].

ПО Docker предназначено для автоматизации развертывания и управления приложениями в средах с поддержкой контейнеризации приложений, а также для более эффективного использования системы и ресурсов, быстрого развертывания программных продуктов, их масштабирования и переноса в другие среды с гарантированным сохранением стабильной работы.

Основной принцип работы Docker – контейнеризация приложений. Этот тип виртуализации позволяет упаковывать программное обеспечение по изолированным средам-контейнерам. Каждая из сред содержит все нужные элементы для работы приложения. Это дает возможность одновременного запуска большого количества контейнеров на одном источнике.

ПО Docker состоит из нескольких компонентов. Во-первых, сервер, выполняющий инициализацию демона (фоновой программы), который применяется для управления и модификации контейнеров, образов и томов. Демон управляет Docker-объектами (сети, хранилища, образы и контейнеры) и может связываться с другими демонами для управления сервисами Docker. Во-вторых, клиент, позволяющий пользователю взаимодействовать с сервером при помощи команд. В-третьих, механизм REST API, отвечающий за организацию взаимодействия Docker-клиента и Docker-демона.

Docker характеризуется достаточно простым синтаксисом, а также совместимо со всеми версиями операционных систем Linux и Windows.

Таким образом, с помощью ПО Docker сервис Калькулятора КТ может быть упакован в контейнер и интегрирован во внешнюю среду. Доступ к такому контейнеру также предоставляется через REST API.

В декабре 2020 года “КТ-калькулятор” был встроен в городскую медицинскую информационную систему, бесплатный доступ к нему получили врачи из любых регионов, а также обычные пользователи.

6. ЗАКЛЮЧЕНИЕ

В данной работе была рассмотрена задача прогнозирования степени поражения легких пациента с диагностированным COVID-19.

Результаты компьютерной томографии являются важным фактором для определения дальнейшей стратегии лечения пациента. При легкой степени поражения КТ 0-1 пациент не требует госпитализации и может проходить лечение на дому, а при тяжелой степени поражения КТ 3-4 пациента небходимо госпитализировать в стационар для проведения интенсивной терапии без промежуточного посещения КТ-центра или поликлиники.

Массовое применение КТ имеет и недостатки, например, риск создания искусственных эпидемических очагов, нерациональная работа бригад скорой помощи и экономические затраты.

В качестве альтернативного диагностического инструмента в рамках данной работы разработаны прогнозные модели оценки легкой (КТ 0-1) и тяжелой (КТ 3-4) степени поражения легких пациента при COVID-19 на основе осмотровых, клинических и биохимических признаков.

Важной особенностью данной работы является использование реальных медицинских данных, имеющих несколько значимых проблем, таких как: сложность сбора данных из нескольких источников, ограниченность сбора данных, несбалансированность заполненности признаков, а также наличие ошибок ввода и противоречий.

В данной работе рассматриваются как базовые модели машинного обучения: метод случайного леса, нейронные сети, градиентный бустинг, так и ансамбли базовых моделей, позволяющие обработать признаки с разной заполненностью. Для сведения откликов прогнозных моделей на вероятностную шкалу используется калибровка Платта.

По результатам экспериментального исследования лучшее качество по метрике ROC AUC показал метод LGBM с заполнением пропусков.

Предложенные прогнозные модели были реализованы в виде веб-сервиса “КТ-калькулятор”, и в декабре 2020 года сервис был встроен в городскую медицинскую информационную систему. Использование сервиса помогает в оперативности принятия врачебных решений и позволяет сократить количество исследований для пациента и

уменьшить облучение. Также, использование сервиса сокращает нагрузку на оборудование КТ-центров и может применяться в регионах, в которых доступ к компьютерному томографу ограничен, либо этого оборудования нет.

СПИСОК ЛИТЕРАТУРЫ

1. *García-Cremades S., Morales-García J., Hernández-Sanjaime R., Martínez-España R., Bueno-Crespo A., Hernáandez-Orallo E., López-Espín J.J., Cecilia J.M.* Improving prediction of Covid-19 evolution by fusing epidemiological and mobility data // Scientific Reports. 2021. V. 11. № 1. P. 1–16.
2. *Elaziz M.A., Hosny K.M., Salah A., Darwish M.M., Lu S., Sahlol A.T.* New machine learning method for image-based diagnosis of covid-19 // Plos One. 2020. V. 15. № 6. P. e0235187.
3. *Levy T.J., Richardson S., Coppa K., Barnaby D.P., McGinn T., Becker L.B., Davidson K.W., Cohen S.L., Hirsch J.S., Zanos T.P. et al.* Development and validation of a survival calculator for hospitalized patients with Covid-19. MedRxiv, 2020.
4. *Jin J., Agarwala N., Kundu P., Harvey B., Zhang Y., Wallace E., Chatterjee N.* Individual and community-level risk for Covid-19 mortality in the united states // Nature Medicine.2021. V. 27. № 2. P. 264–269.
5. *Yadaw A.S., Li Y.-C., Bose S., Iyengar R., Bunyavanich S., Pandey G.* Clinical features of Covid-19 mortality: development and validation of a clinical prediction model // Lancet Digital Health. 2020. V. 2. № 10. P. e516–e525.
6. *Shah S., Majmudar K., Stein A., Gupta N., Suppes S., Karamanis M., Capannari J., Sethi S., Patte C.* Novel use of home pulse oximetry monitoring in Covid-19 patients discharged from the emergency department identifies need for hospitalization // Academic Emergency Medicine. 2020. V. 27. № 8. P. 681–692.
7. *Bishop C.M.* Pattern recognition // Machine Learning. 2006. V. 128. № 9.
8. *Ripley B.D.* Pattern recognition and neural networks. Cambridge University Press, 2007.
9. *Breiman L.* Random forests // Machine Learning. 2001. V. 45. № 1. P. 5–32.
10. *Breiman L., Friedman J.H., Olshen R.A., Stone C.J.* Classification and regression trees. Routledge, 2017.
11. *Rumelhart D.E., Hinton G.E., Williams R.J.* Learning representations by backpropagating errors // Nature. 1986. V. 323. № 6088. P. 533–536.
12. *Minsky M.L., Papert S.A.* Perceptrons: expanded edition, 1988.
13. *Hunt K.J., Sbarbaro D., Żbikowski R., Gawthrop P.J.* Neural networks for control systems – a survey // Automatica. 1992. V. 28. № 6. P. 1083–1112.
14. *Breiman L.* Randomizing outputs to increase prediction accuracy // Machine Learning. 2000. V. 40. № 3. P. 229–242.
15. *Friedman J.H.* Greedy function approximation: a gradient boosting machine // Annals of Statistics. 2001. P. 1189–1232.
16. *Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T.-Y.* Lightgbm: A highly efficient gradient boosting decision tree // Advances in Neural Information Processing Systems. 2017. V. 30. P. 3146–3154.
17. *Dreiseitl S., Ohno-Machado L.* Logistic regression and artificial neural network classification models: a methodology review // Journal of Biomedical Informatics. 2002. V. 35. № 5–6. P. 352–359.
18. *Rosen B.E.* Ensemble learning using decorrelated neural networks // Connection Science. 1996. V. 8. № 3–4. P. 373–384.
19. *DeGroot M.H., Fienberg S.E.* The comparison and evaluation of forecasters // Journal of the Royal Statistical Society: Series D (The Statistician). 1983. V. 32. № 1–2. P. 12–22.
20. *Niculescu-Mizil A., Caruana R.* Obtaining calibrated probabilities from boosting / In: UAI. 2005. V. 5. P. 413–420.
21. *Platt J. et al.* Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods // Advances in Large Margin Classifiers. 1999. V. 10. № 3. P. 61–74.
22. *Hand D.J., Till R.J.* A simple generalisation of the area under the roc curve for multiple class classification problems // Machine Learning. 2001. V. 45. № 2. P. 171–186.
23. *Smith G.B., Redfern O.C., Pimentel M.A., Gerry S., Collins G.S., Malycha J., Prytherch D., Schmidt P.E., Watkinson P.J.* The national early warning score 2 (news2) // Clinical Medicine. Journal of the Royal College of Physicians of London. 2019. V. 19. № 3.
24. *Gardner J.* The web server gateway interface (wsgi) // The Definitive Guide to Pylons. 2009. P. 369–388.
25. *Ong S.P., Cholia S., Jain A., Brafman M., Gunter D., Ceder G., Persson K.A.* The materials application programming interface (api): A simple, flexible and efficient api for materials data based on representational state transfer (rest) principles // Computational Materials Science. 2015. V. 97. P. 209–215.
26. *Crockford D.* The application/json media type for javascript object notation (json) // RFC 4627, 2006.
27. *Anderson C.* Docker [software engineering] // IEEE Software. 2015. V. 32. № 3. P. 102_c3.
28. *Boettiger C.* An introduction to docker for reproducible research // ACM SIGOPS Operating Systems Review. 2015. V. 49. № 1. P. 71–79.