УДК 004.8
DOI: 10.26102/2310-6018/2023.42.3.014

Разработка методов прогнозирования динамики заболеваемости на примере COVID-19

И.Л. Каширина✉, О.В. Матыкина

Воронежский государственный университет, Воронеж, Российская Федерация

Резюме. Пандемия COVID-19 привела к глобальным последствиям и стала причиной серьезных ограничительных мер во всех сферах деятельности, изменивших условия работы и жизни населения мира. Даже после окончания пандемии прогнозирование заболеваемости COVID-19 остается важной задачей, так как необходимо следить за развитием ситуации, а результаты исследований по этой теме могут быть перенесены на другие эпидемии. Особое значение имеют научные исследования по анализу факторов, оказывающих существенное влияние на протекание эпидемии. В данном исследовании предлагается комплекс моделей и алгоритмов машинного обучения, базирующихся на обработке больших данных, для прогнозирования динамики распространения вируса COVID-19 на мезоуровне, с помощью которого анализируется влияние различных экзогенных факторов на заболеваемость. В качестве исходных данных для построения моделей машинного обучения используется деперсонифицированный набор данных, предоставленный Воронежским областным клиническим консультативно-диагностическим центром и содержащий информацию по всем проведенным в Воронежской области тестам на COVID-19. Для эффективной борьбы с эпидемиями необходимы прогнозы развития динамики заболеваемости на достаточно длительный период времени (например, от двух недель и более), тогда как в литературе, как правило, предлагаются краткосрочные методы, позволяющие делать достаточно точный прогноз только на 1–5 дней. Поэтому задача данного исследования заключается в поиске оптимального метода прогнозирования заболеваемости на средний период времени с использованием экзогенных факторов. В качестве экзогенных переменных для улучшения качества прогнозирования были выбраны сведения о погоде, дне недели и месяце и популярность поисковых запросов, связанных с COVID-19.

Ключевые слова: COVID-19, машинное обучение, временные ряды, прогнозирование динамики, гибридная нейронная сеть.

Для цитирования: Каширина И.Л., Матыкина О.В. Разработка методов прогнозирования динамики заболеваемости на примере COVID-19. *Моделирование, оптимизация и информационные технологии.* 2023;11(3). URL: https://moitvivt.ru/ru/journal/pdf?id=1434 DOI: 10.26102/2310-6018/2023.42.3.014

Development of methods for forecasting the dynamics of morbidity in the case of COVID-19

I.L. Kashirina✉, O.V. Matykina

Voronezh State University, Voronezh, the Russian Federation

Abstract. The COVID-19 pandemic has had global repercussions and has led to severe restrictive measures in all areas of activity that have changed the working and living conditions of the world's population. Even after the end of the pandemic, predicting the incidence of COVID-19 remains an important task as it is necessary to monitor the development of the situation and the results of research on this issue can be extrapolated to other epidemics. Scientific studies on the analysis of factors that have a significant impact on the course of the epidemic have a particular importance. This study proposes a set of models and machine learning algorithms based on big data processing to predict the dynamics of the spread of the COVID-19 virus at the mesolevel, which analyzes the impact of various

exogenous factors on the incidence. As the initial data for building machine learning models, we use a depersonalized data set provided by Voronezh Regional Clinical Consultative and Diagnostic Center and containing information on all tests for COVID-19 conducted in Voronezh Oblast. To effectively combat epidemics, it is necessary to forecast the development of the incidence dynamics for a sufficiently long period of time, e.g. from two weeks or more, while various studies, in general, propose short-term methods that allow making a fairly accurate forecast only for 1–5 days. Therefore, the goal of this study is to find the optimal method for predicting incidence over an average period of time using exogenous factors. Information about the weather, day of the week and month, and the popularity of search queries related to COVID-19 were selected as exogenous variables to improve the quality of forecasting.

Keywords: COVID-19, machine learning, time series, dynamics prediction, hybrid neural network.

For citation: Kashirina I.L., Matykina O.V. Development of methods for forecasting the dynamics of morbidity in the case of COVID-19. *Modeling, Optimization and Information Technology*. 2023;11(3). URL: https://moitvivt.ru/ru/journal/pdf?id=1434 DOI: 10.26102/2310-6018/2023.42.3.014 (In Russ.).

Введение

Пандемия COVID-19 стала серьезной проблемой, имеющей международное значение. На этом фоне широко востребованными стали технологии прогнозной аналитики, активно использующиеся для составления сценариев распространения эпидемии заболеваемости. Даже после окончания пандемии прогнозирование развития COVID-19 всё ещё является важной задачей. Для принятия правильных решений нужна достоверная информация об эпидемиологической ситуации и ее прогнозах. В последнее время очень популярными при моделировании эпидемиологических процессов стали алгоритмы машинного обучения, ключевой особенностью которых является способность самостоятельно выделять закономерности в данных на основе имеющейся базы наблюдений.

Существует множество методов прогнозирования, но для эффективной борьбы с эпидемиями необходимы точные прогнозы на достаточно длительный период времени. Хотя нейросетевые модели показывают отличные результаты в краткосрочных прогнозах [1, 2], для задачи прогнозирования динамики заболеваемости горизонт предсказания должен быть больше, чтобы система здравоохранения имела возможность перераспределить ресурсы. Среднесрочные методы, как правило, имеют более низкую точность, чем краткосрочные. Поэтому задача заключается в поиске оптимального метода прогнозирования заболеваемости на средний период времени с высокой точностью. Целью исследования является разработка моделей среднесрочного прогнозирования динамики заболеваемости COVID-19 в Воронежской области и проведение их сравнительного анализа.

В проведенных ранее исследованиях по данной теме [3, 4] представлен разведочный анализ исходных данных, и реализована интерактивная визуализация динамики распространения COVID-19 на карте региона в Google, а также описана начальная (baseline) модель прогнозирования, базирующаяся на рекуррентных нейронных сетях. В данной работе важный упор сделан на выявление дополнительных признаков, улучшающих точность прогнозирования, и на поиск методов, наиболее точных для построения ежедневных прогнозов на 20 дней вперед.

В качестве дополнительных признаков, в частности, рассматривается статистика поисковых запросов, связанных с COVID-19. Данный подход показал свою эффективность в нескольких международных исследованиях. Например, в [5, 6] авторы предложили использовать модели прогнозирования заболеваемости COVID-19 с использованием данных Google Trends. Статистика поисковых запросов Google является открытым и доступным источником. В [5] показано, что статистика запросов по словам

«coronavirus» и «антисептик» позволяет повысить точность прогнозирования распространения вируса в Иране. В [6] было обнаружено, что максимальный период задержки для прогнозирования случаев COVID-19 в Индии приходится на поисковые запросы по термину «коронавирус» и составляет 21 день, т. е. число поисковых запросов по слову «коронавирус» имеет максимальную корреляцию с числом случаев COVID-19, зарегистрированных системой эпиднадзора за заболеваниями через 21 день.

В целом поиску дополнительных факторов, влияющих на распространение вируса, посвящено довольно много научных публикаций. Например, в [7] в качестве экзогенных переменных рассматриваются показатели качества воздуха и автомобильный траффик. В данном исследовании рассматривается возможность использования метеорологических данных из открытых источников при прогнозировании эпидемии.

Для решения задачи прогнозирования в данном исследовании предлагается нейросетевая модель гибридной архитектуры, включающая сверточные и рекуррентные слои. Предлагаемая модель сравнивается с существующими подходами, чтобы выявить ее преимущества и недостатки.

Исходные данные

Построение моделей машинного обучения происходило на наборе обезличенных данных, предоставленных Воронежским областным клиническим консультативно-диагностическим центром (ВОККДЦ). Набор включает в себя данные о результатах всех ПЦР-тестов на COVID-19, которые проводились в Воронежской области в промежутке с марта 2020 года по октябрь 2022 года. Фрагмент датасета представлен на Рисунке 1. Датасет содержит около 3 миллионов записей о результатах ПЦР-тестирования.

ГУИД	Пол	Год рождения	Дата забора	Результат	МО по месту жительства	Направлен	Стационар тяжёлый	Осложнения после ковид	Жив	Вторая вакцина более двух недель назад
ee00dbf8-	Мужской	2008	27.10.2022	Отр.	БУЗ ВО "Рамонская РБ"	Амбулаторно	Нет	Нет	Да	Нет
ee00dc0	Женский	1963	27.10.2022	Отр.	БУЗ ВО "Новоусманская	Амбулаторно	Нет	Нет	Да	Нет
ee00dc0	Мужской	1974	27.10.2022	Пол.	БУЗ ВО "Семилукская РБ	Амбулаторно	Нет	Нет	Да	Да
5ade159	Мужской	2009	27.10.2022	Отр.	БУЗ ВО "Павловская РБ"	Амбулаторно	Нет	Нет	Да	Нет
5ade15a	Женский	2015	27.10.2022	Отр.	БУЗ ВО "Новоусманская	Амбулаторно	Нет	Нет	Да	Нет
5ade15a	Мужской	2020	27.10.2022	Отр.	БУЗ ВО "Павловская РБ"	Амбулаторно	Нет	Нет	Да	Нет
873b4e4	Женский	2019	27.10.2022	Отр.	БУЗ ВО "ВГКБ №11"	Амбулаторно	Нет	Нет	Да	Нет
0657e0b	Мужской	1984	27.10.2022	Пол.	БУЗ ВО "Бобровская РБ"	Стационарно	Нет	Нет	Да	Нет
0657e0b	Мужской	2002	27.10.2022	Отр.	БУЗ ВО "ВГКП № 1"	Амбулаторно	Нет	Нет	Да	Да
3be3e7e	Женский	2007	27.10.2022	Отр	БУЗ ВО "Семилукская РБ	Стационарно	Нет	Нет	Да	Нет
4c4d2b5	Женский	2014	27.10.2022	Отр.	БУЗ ВО "ВГКБ №11"	Амбулаторно	Нет	Нет	Да	Нет
4c4d2b5	Женский	1965	27.10.2022	Отр.	БУЗ ВО "ВГКП № 7"	Амбулаторно	Нет	Нет	Да	Нет
4c4d2b5	Женский	2004	27.10.2022	Отр.	БУЗ ВО "ВГКП № 1"	Амбулаторно	Нет	Нет	Да	Нет

Рисунок 1 − Фрагмент исходных данных
Figure 1 − Fragment of initial data

Первоначально была выдвинута гипотеза о том, что добавление в качестве экзогенных признаков погодных показателей может улучшить прогноз. Для того чтобы проверить эту гипотезу, был собран датасет с данными о погоде за каждые сутки в рассматриваемый период. Датасет метеорологических наблюдений был загружен с сайта «Расписание Погоды», rp5.ru. В набор данных входят следующие измерения: средняя температура воздуха за сутки (T), максимальная по модулю температура воздуха за сутки (T_abs_max), среднее атмосферное давление на уровне станции (Po), изменение атмосферного давления между за сутки (Pa), средняя относительная влажность в сутки (U_mean), суточное количество выпавших осадков (RRR).

К полученному набору был добавлен столбец количества заболевших (Id), и была построена матрица корреляции. Затем были отобраны те сведения о погоде, модуль

коэффициента корреляции целевого столбца с которыми больше или равен 0,2. Этими признаками оказались среднесуточная относительная влажность и максимальное по модулю значение температуры за сутки. Матрица корелляции отобранных признаков и целевого столбца представлена на Рисунке 2. Отобранные метеорологические показатели были добавлены к ряду количества заболевших.

	id	U_mean	T_abs_max
id	1.000000	0.226055	-0.325035
U_mean	0.226055	1.000000	-0.614093
T_abs_max	-0.325035	-0.614093	1.000000

Рисунок 2 − Матрица корреляции сведений о погоде и количества заболевших
Figure 2 − Covariance matrix of weather information and the number of cases

Далее была выдвинута гипотеза о том, что добавление к модели ряда количества поисковых запросов может улучшить качество прогнозирования. Информация о динамике запросов собиралась с помощью Google Trends. Google Trends – это инструмент, который показывает статистику запросов пользователей в поисковой системе Google. Полученные данные анонимизируются, сгруппировываются по темам и объединяются в группы. Информацию можно просмотреть по всем странам и городам, где используется Google.

Числа, которые отображаются на графике, указывают на уровень интереса к теме в сравнении с наивысшим показателем в данном регионе и периоде времени. Например, 100 баллов означают наивысший уровень популярности запроса, а 50 баллов – уровень популярности в два раза меньший. Если о запросе нет достаточной информации, то ему присваивается 0 баллов.

На Рисунке 3 представлены ряд количества заболевших (вверху) и график динамики популярности запроса «COVID-19» (внизу) в Воронежской области в рассматриваемый период. Видно, что график заболеваемости похож на график поисковых запросов, сдвинутый вперед. Именно с запросом «COVID-19» в Воронежской области обнаружилась самая высокая корреляция с лагом в 21 день (0.37), что согласуется с результатами, представленными в [6]. Также были проанализированы запросы по словам «коронавирус», «лечение коронавируса», «COVID», но с ними корреляция оказалась ниже.

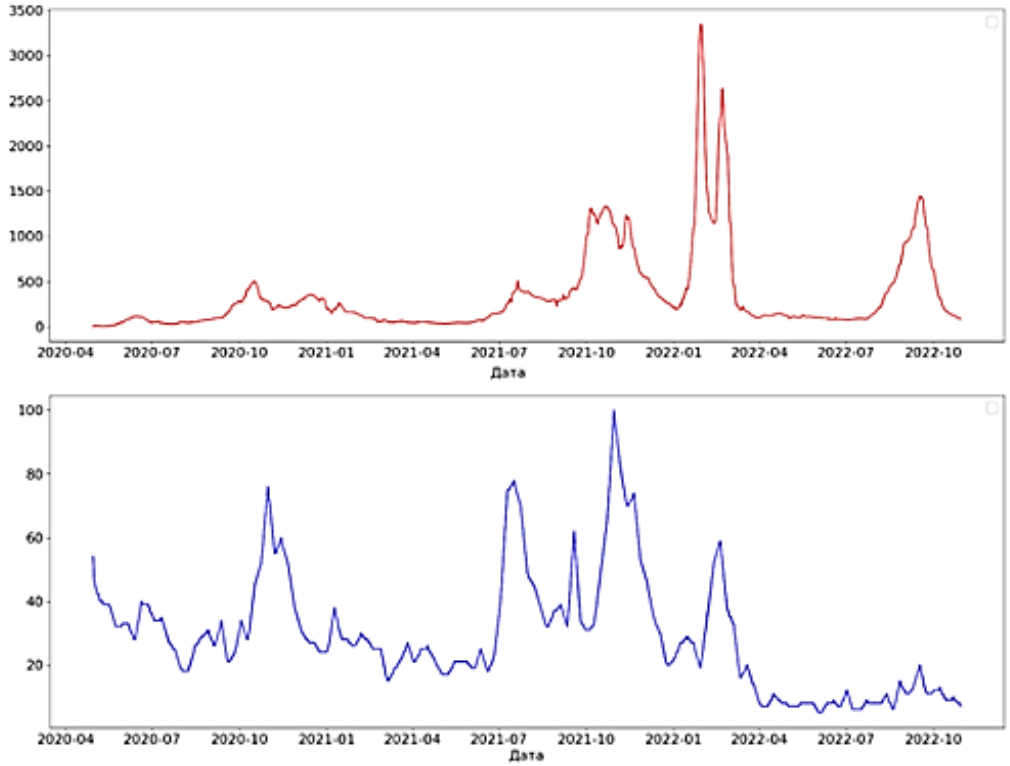

Рисунок 3 − График ряда количества заболевших и график популярности запроса «COVID-19» в Воронежской области

Figure 3 − Graphs of the number of coronavirus cases and the popularity of COVID-19 search query in Voronezh Oblast

Материалы и методы

В качестве основной модели для прогнозирования динамики заболеваемости на 21 день вперед была выбрана гибридная нейросетевая архитектура CNN-LSTM, которая ранее хорошо себя зарекомендовала в исследованиях [8, 9].

Традиционно разработанные для обработки изображений сверточные нейронные сети CNN [10] могут использоваться и для прогнозирования временных рядов. Разница заключается в формате входных данных. Вместо двумерных матриц изображений на вход модели поступает вектор наблюдений временного ряда (или несколько временных рядов, если дополнительно используются ряды с экзогенными признаками). Ядра имеют ширину k, равную количеству одновременно подаваемых на вход признаков, но могут иметь различную длину. Ядро перемещается по многомерному временному ряду в одном направлении, выполняя свертку. В отличие от двумерной свертки изображений, ядро не перемещается влево или вправо. Элементы ядра умножаются на соответствующие элементы временного ряда и результаты складываются, после чего к ним применяется нелинейная функция активации. Полученное значение становится элементом нового «отфильтрованного» временного ряда, а ядро продвигается дальше по временному ряду для получения следующего значения. Количество новых «отфильтрованных» временных рядов соответствует количеству ядер свертки. Различные свойства и особенности исходного временного ряда отражаются в новых фильтрованных рядах. Иллюстрация CNN для обработки временных рядов представлена на Рисунке 4.

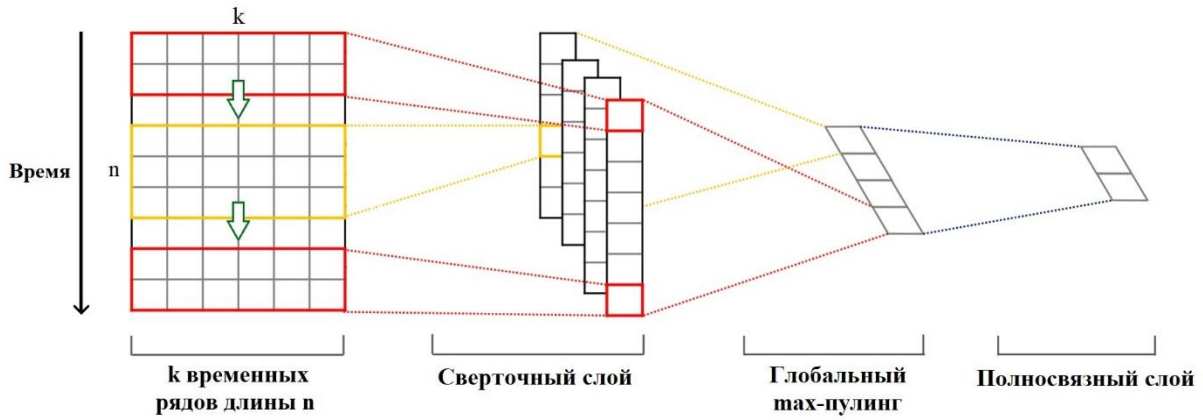

Рисунок 4 − CNN для временного ряда
Figure 4 − CNN for a time series

В отличие от стандартных сетей прямого распространения, LSTM-сеть имеет обратные соединения. Такая рекуррентная нейронная сеть может успешно обрабатывать не только отдельные точки, но и целые последовательности данных, что характерно для задач прогнозирования временных рядов. Особенностью рекуррентной архитектуры LSTM является то, что в процессе обучения такой сети моделируется не только долговременная (Long), но и кратковременная (Short) память [11]. Базовая архитектура LSTM включает четыре слоя. Слой g_t является основным слоем, анализирующим поток входных данных и предыдущее состояние краткосрочной памяти. Остальные три слоя являются контроллерами шлюзов (gate controller). Шлюз забывания (f_t) управляет тем, какие части долгосрочной памяти (h_t) должны быть удалены. Входной шлюз (i_t) управляет тем, какие (наиболее значимые) части g_t должны быть сохранены в долгосрочной памяти. Выходной шлюз (o_t) управляет тем, какие части долгосрочного состояния следует выдавать в качестве результата. После прохождения всех шлюзов вычисляется текущее выходное состояние ячейки LSTM.

Идея совместного применения моделей CNN и LSTM для решения задачи прогнозирования временных рядов состоит в том, что сверточные слои будут извлекать значимые признаки из заданного набора данных временных рядов, а слой LSTM идентифицировать долгосрочные и краткосрочные зависимости. Использованная в данном исследовании архитектура сети CNN-LSTM представлена на Рисунке 5.

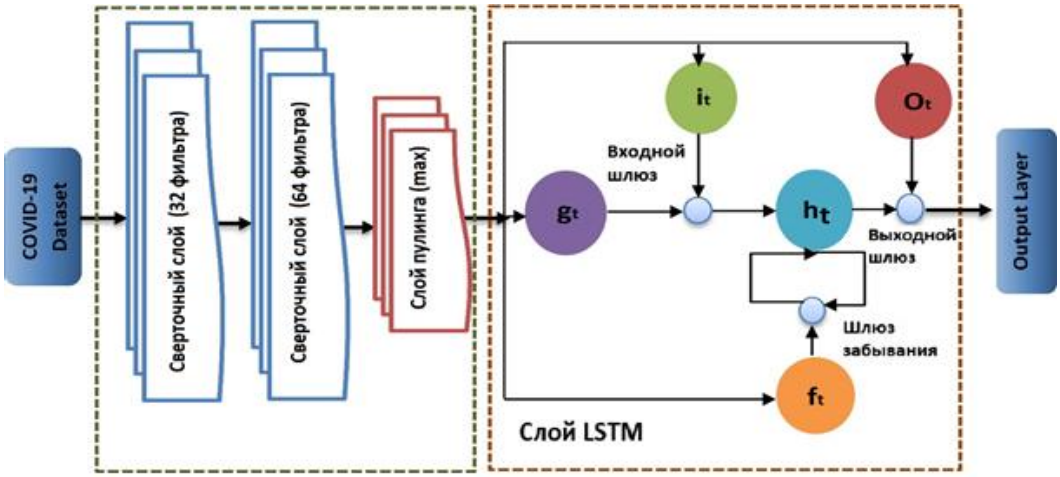

Рисунок 5 − Архитектура сети CNN-LSTM
Figure 5 − CNN-LSTM network architecture

Помимо описанной гибридной архитектуры в данном исследовании было протестировано еще несколько моделей для решения поставленной задачи. В частности, чтобы оценить возможности автоматического машинного обучения, была использована библиотека AutoTS. AutoTS – это набор инструментов на языке Python, предназначенный для быстрого и точного прогнозирования временных рядов. В 2023 году модель, построенная в библиотеке AutoTS, выиграла ежегодный (проводимый с 1982 года) международный конкурс прогнозирования временных рядов M6. В пакет AutoTS включены десятки моделей прогнозирования, такие как статистические, машинного и глубокого обучения, а также более 30 преобразований, специфичных для временных рядов. Все модели работают напрямую с фреймами данных Pandas без необходимости преобразования в другие объекты. AutoTS поддерживает прогнозирование многомерных выходных данных и вероятностные прогнозы. Большинство моделей поддерживают передачу определяемых пользователем экзогенных признаков. В зависимости от ограничений по времени можно выбрать подмножество моделей, а не обучать последовательно все.

Так как в нескольких источниках [5, 6] было отмечено, что высокую точность предсказания динамики COVID-19 показывают статистические модели, в данном исследовании также использовалась модель SARIMAX. SARIMAX (Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors) – это обновленная версия модели ARIMA. ARIMA включает в себя авторегрессионную интегрированную скользящую среднюю, SARIMAX включает, помимо этого, сезонные эффекты и экзогенные факторы. Модель SARIMAX имеет четыре параметра (p, d, q, s), где p – порядок авторегрессионной модели (количество используемых для прогнозирования предшествующих временных лагов); d – порядок разности (количество раз, когда из данных вычитались прошлые значения); q – порядок в модели скользящей средней; s – размер сезонности.

Результаты и обсуждение

В качестве метрик точности прогноза использовались коэффициент детерминации R^2, корень из средней квадратичной ошибки $RMSE$ и средняя абсолютная ошибка в процентах $MAPE$, которая показывает, на сколько процентов в среднем спрогнозированные значения отличались от реальных.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - f(x_i))^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2},\ RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - f(x_i)\right)^2},\ MAPE = \frac{1}{N}\frac{\sum_{i=1}^{N}|y_i - f(x_i)|}{|y_i|} \quad (1)$$

В формулах (1) $y_i -$ реальное, $f(x_i) -$ предсказанное, $\overline{y} -$ среднее значение временного ряда. Все метрики рассчитывались для пошагового ежедневного прогноза на 21 день вперед.

Первые модели были построены без использования экзогенных переменных. Прогноз строился, опираясь только на предшествующие значения ряда ежедневной динамики числа заболевших в Воронежской области. В качестве горизонта прогноза установлено значение 21 день.

На Рисунке 6 представлены графики реального и предсказанного количества заболевших, полученные с помощью автоматического подбора модели и параметров библиотеки автоматического прогнозирования временных рядов AutoTS.

Далее, в качестве базовой модели была взята SARIMAX. Путем подбора параметров были найдены такие, при которых метрики показывали лучший результат. Такими параметрами оказались (1, 0, 0, 2).

В качестве основного подхода разработана гибридная модель глубокого обучения

CNN-LSTM. На Рисунке 8 представлены графики реального и предсказанного количества заболевших.

Значения метрик представлены в Таблице 1.

Таблица 1 − Метрики на данных «ряд количества заболевших» без экзогенных переменных
Table 1 – Metrics on the number of coronavirus cases not accounting for exogenous variables

Модель	R^2	RMSE	MAPE
AutoTS	0.785237	18.574383	10.141119
SARIMAX	0.909123	12.082673	8.516641
CNN-LSTM	0.915886	11.624374	7.072621

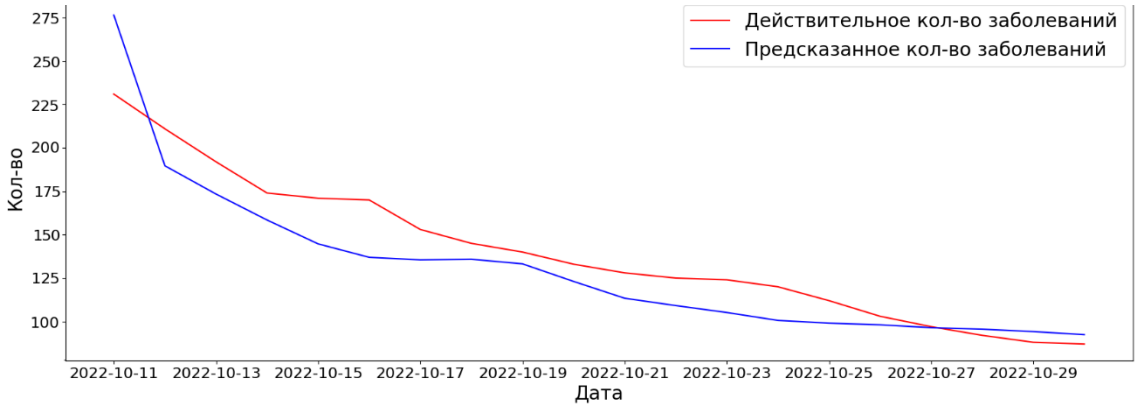

Рисунок 6 − Прогнозирование динамики заболеваемости с помощью AutoTS
Figure 6 − Forecasting morbidity dynamics using AutoTS

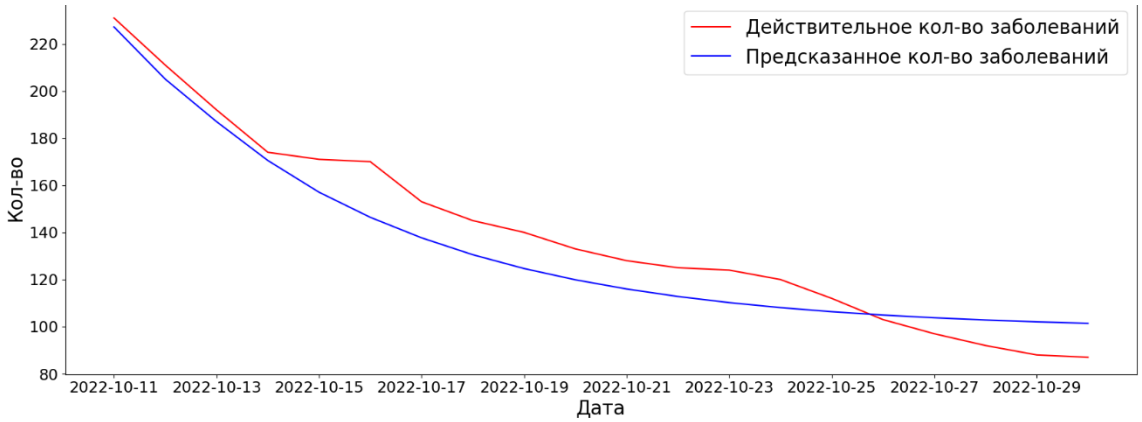

Рисунок 7 − Прогнозирование динамики заболеваемости с помощью SARIMAX
Figure 7 − Forecasting morbidity dynamics using SARIMAX

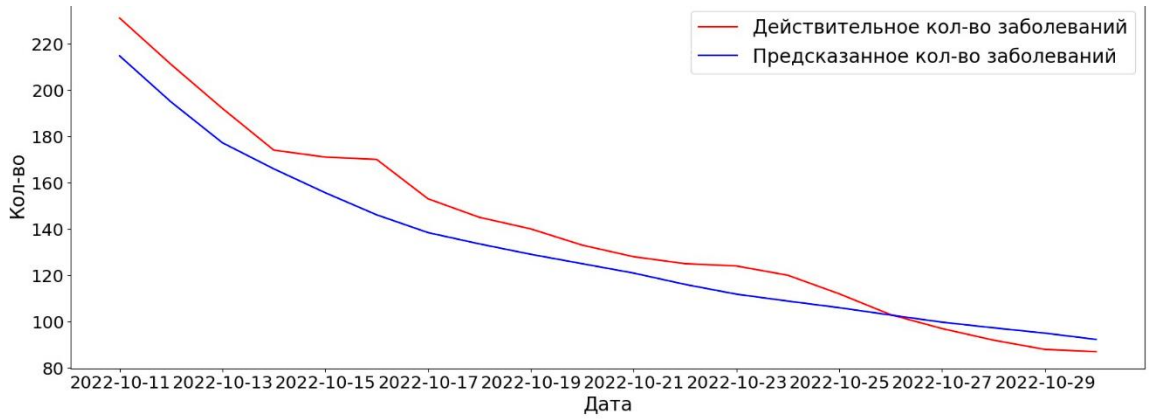

Рисунок 8 − Прогнозирование динамики заболеваемости с помощью CNN-LSTM
Figure 8 − Forecasting morbidity dynamics using CNN-LSTM

Как видно из результатов, при обучении на данных без добавления экзогенных признаков лучшей оказалась модель CNN-LSTM, худшей − модель AutoTS. Так как обучение модели AutoTS занимает очень много времени, далее она не тестировалась.

На следующем этапе датасет был дополнен описанными ранее экзогенными признаками. Первой была обучена модель SARIMAX. Параметры использовались те же, что и в модели SARIMAX без экзогенных переменных. Первоначально в качестве экзогенных переменных выборки брались реальные данные о погоде за соответствующий период (без лага). Такой прогноз является однодневным. На Рисунке 9 представлены графики реального и предсказанного количества заболевших.

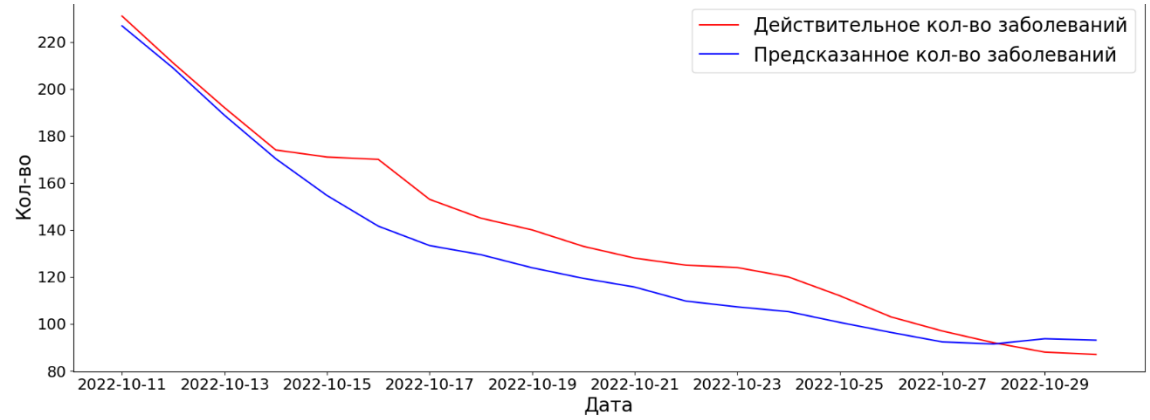

Рисунок 9 − Прогнозирование динамики заболеваемости с помощью SARIMAX с экзогенными переменными погоды
Figure 9 − Forecasting morbidity dynamics using SARIMAX with exogenous weather variables

Аналогичный прогноз был построен с помощью гибридной CNN-LSTM модели. На Рисунке 10 представлены графики реального и предсказанного этой моделью количества заболевших.

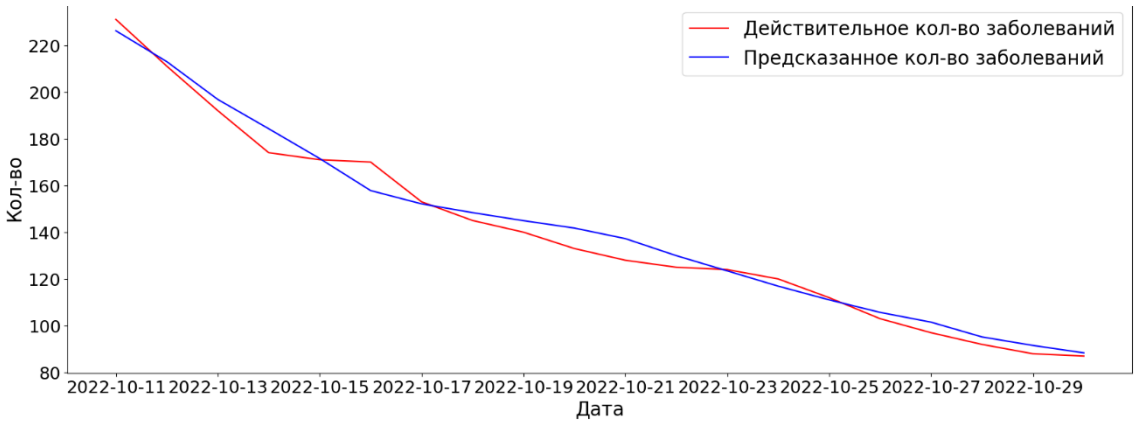

Рисунок 10 − Прогнозирование динамики заболеваемости с помощью CNN-LSTM с экзогенными переменными погоды
Figure 10 − Forecasting morbidity dynamics using CNN-LSTM with exogenous weather variables

Метрики одношагового прогноза моделей на данных «ряд количества заболевших» с погодой представлены в Таблице. 2. Из Таблицы 2 видно, что точность модели SARIMAX упала (возможно, потому что после добавления погодных данных потребовался подбор новых параметров), а точность CNN-LSTM существенно возросла.

Таблица 2 − Метрики одношагового прогноза SARIMAX и CNN-LSTM на данных «ряд количества заболевших» с погодой
Table 2 – Metrics of SARIMAX and CNN-LSTM one-step forecast using the data on the number of coronavirus cases accounting for weather information

Модель	R^2	RMSE	MAPE
SARIMAX	0.89599	12.926259	8.066269
CNN-LSTM	0.981623	5.433444	3.164966

Далее было проведено обучение моделей на данных, где ряды погоды сдвинуты на 21 день вперед. Это позволило делать предсказания сразу на 21 день вперед (многошаговый прогноз). Результат работы SARIMAX представлен на Рисунке 11, а модели CNN-LSTM на Рисунке 12.

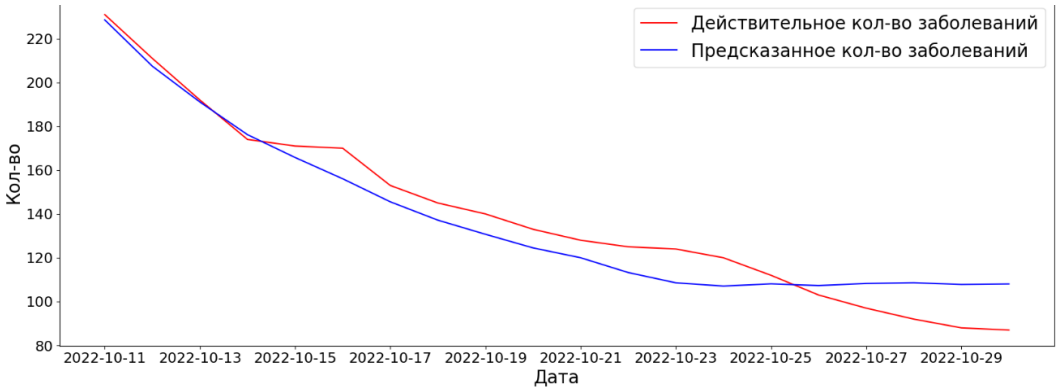

Рисунок 11 − Прогнозирование динамики заболеваемости с помощью SARIMAX с экзогенными переменными погоды со сдвигом 21 день
Figure 11 − Forecasting morbidity dynamics using SARIMAX with exogenous weather variables with 21-day lag

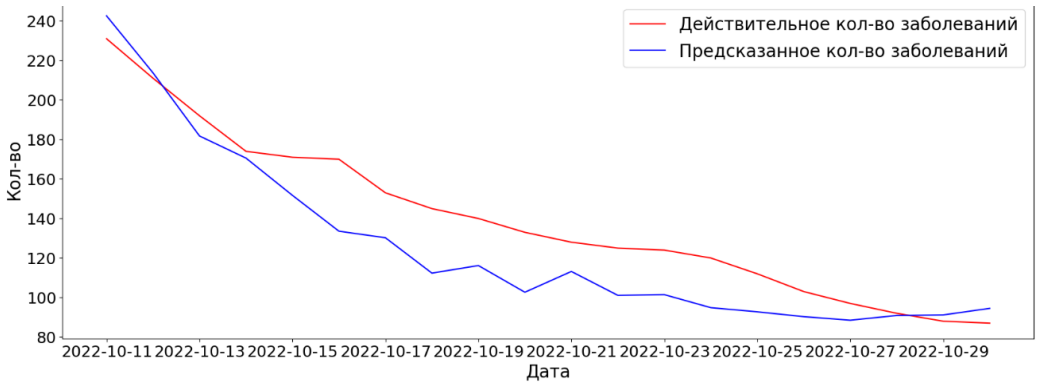

Рисунок 12 − Прогнозирование динамики заболеваемости с помощью CNN-LSTM с экзогенными переменными погоды со сдвигом 21 день
Figure 12 − Forecasting morbidity dynamics using CNN-LSTM with exogenous weather variables with a 21-day lag

Метрики результата работы моделей на данных «ряд количества заболевших» с погодой представлены в Таблице 3. Из Таблиц 2 и 3 можно сделать вывод о том, что добавление сведений о погоде улучшает прогноз, причем результат CNN-LSTM значительно улучшился при одношаговом прогнозировании, а прогноз SARIMAX стал точнее на многодневном предсказании.

Таблица 3 − Метрики SARIMAX и CNN-LSTM на данных «ряд количества заболевших» с погодой с лагом в 21 день
Table 3 − SARIMAX and CNN-LSTM metrics using the data on the number of coronavirus cases with a 21-day lag

Модель	R^2	RMSE	MAPE
SARIMAX	0.925317	10.953356	8.094839
CNN-LSTM	0.862394	12.537271	9.254269

Затем на датасете, полученном объединением ряда количества заболевших с рядом относительной популярности запроса, была обучена модель SARIMAX с экзогенными признаками. Параметры использовались такие же, как в предыдущих моделях. На Рисунке 13 представлены графики реального и предсказанного количества заболевших.

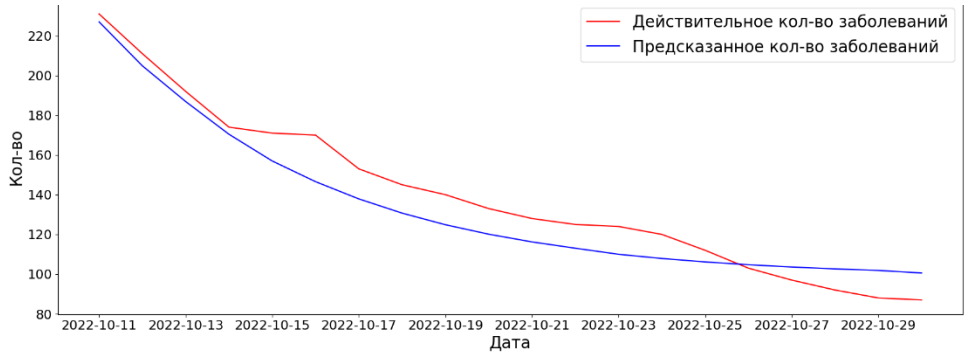

Рисунок 13− Прогнозирование динамики заболеваемости с помощью SARIMAX с экзогенной переменной «популярность запроса» со сдвигом 20 дней
Figure 13 − Forecasting morbidity dynamics using SARIMAX with an exogenous variable of search query popularity with a 21-day lag

На этих данных также была обучена модель CNN-LSTM. На Рисунке 14 представлены графики реального и предсказанного этой моделью количества заболевших.

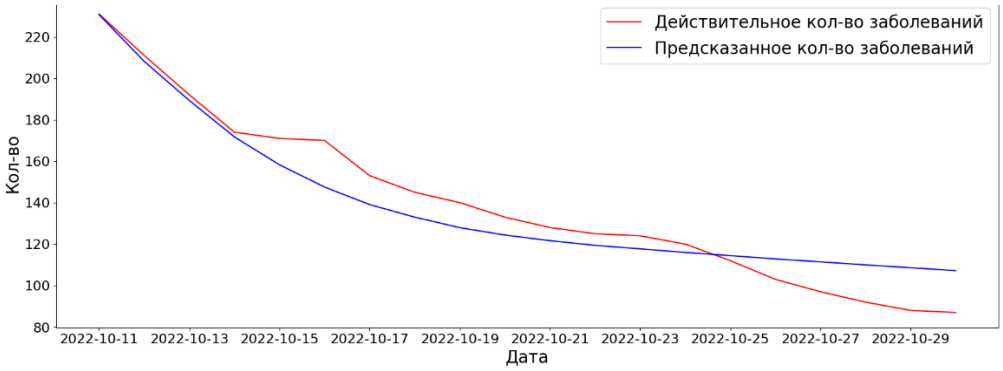

Рисунок 14 − Прогнозирование динамики заболеваемости с помощью CNN-LSTM с экзогенной переменной «популярность запроса» со сдвигом 21 день
Figure 14 − Forecasting morbidity dynamics using CNN-LSTM with an exogenous variable of search query popularity with a 21-day lag

Метрики результата работы моделей на данных «ряд количества заболевших» с популярностью запроса представлены в Таблице 4.

Таблица 4 − Метрики SARIMAX и CNN-LSTM на данных «ряд количества заболевших» с популярностью запроса
Table 4 − SARIMAX and CNN-LSTM metrics using the data on the number of coronavirus cases accounting for search query popularity

Модель	R^2	RMSE	MAPE
SARIMAX	0.920926	11.562161	8.41726
CNN-LSTM	0.921627	11.515006	7.059072

Добавление в качестве экзогенных переменных дня недели и месяца привело к небольшому повышению точности прогноза модели SARIMAX, но ухудшило точность модели CNN-LSTM (результаты представлены в итоговой таблице).

Заключение

Метрики полученных результатов прогнозирования построенных моделей приведены в Таблице 5. Для каждого набора входных данных в таблице выделена модель, показавшая лучший результат по метрикам.

Таблица 5 − Метрики построенных моделей
Table 5 – Metrics of the designed models

Входные данные	Модель	R^2	RMSE	MAPE
Ряд количества заболевших (прогноз на 21 день)	AutoTS	0.785237	18.574383	10.141119
	SARIMAX	0.909123	12.082673	8.516641
	CNN-LSTM	**0.915886**	**11.624374**	**7.072621**

Таблица 5 (продолжение)
Table 5 (extended)

Входные данные	**Модель**	**R^2**	**RMSE**	**MAPE**
Ряды количества заболевших и погоды (одношаговый прогноз)	SARIMAX	0.89599	12.926259	8.066269
	CNN-LSTM	**0.981623**	**5.433444**	**3.164966**
Ряды количества заболевших и погоды (прогноз на 21 день)	**SARIMAX**	**0.925317**	**10.953356**	**8.094839**
	CNN-LSTM	0.862394	12.537271	9.254269
Ряды количества заболевших, дней и месяцев (прогноз на 21 день)	**SARIMAX**	**0.910144**	**12.014598**	**8.577228**
	CNN-LSTM	0.578236	15.817493	19.673749
Ряды количества заболевших и запроса (прогноз на 21 день)	SARIMAX	0.920926	11.562161	8.41726
	CNN-LSTM	**0.921627**	**11.515006**	**7.059072**

Из таблицы видно, что добавление в качестве экзогенных переменных показателей погоды и популярности поискового запроса повышает качество многошагового прогнозирования. При этом лучшее качество прогноза достигнуто на данных, состоящих из ряда количества заболевших с экзогенными признаками «сведения о погоде». Также можно сделать вывод о том, что гибридная CNN-LSTM модель гораздо лучше справилась на одношаговом прогнозе, а модель SARIMAX чуть точнее предсказывала количество заболевших на 21 день вперед. В случае эпидемий прогноз на несколько дней является более полезным, так как позволяет перераспределить ресурсы.

Таким образом, в ходе данного исследования была построена модель машинного обучения на основе нейронных сетей для прогнозирования количества заболевших, проведено ее сравнение со статистической моделью SARIMAX и библиотекой автоматического машинного обучения AutoTS. Для обучения моделей использовались данные о количестве заболевших COVID-19 и дополнительные признаки, такие как сведения о погоде, дне и месяце, популярность поискового запроса. При этом были проверены гипотезы об улучшении качества прогнозирования количества заболевших при добавлении различных экзогенных признаков к моделям.

Результаты оценки качества моделей показывают, что использование дополнительных данных позволяет улучшить точность предсказаний. Разработанные модели в дальнейшем могут использоваться для прогнозирования развития других эпидемий.

СПИСОК ИСТОЧНИКОВ

1. Braga M.D.B, Fernandes R.D.S, de Souza G.N., et al. Artificial neural networks for short-term forecasting of cases, deaths, and hospital beds occupancy in the COVID-19 pandemic at the Brazilian Amazon. *PLoS One.* 2021;16(3):e0248161. DOI: 10.1371/journal.pone.0248161.
2. Naumov A., Moloshnikov I., Serenko A., Sboev A., Rybka R. Baseline accuracies of forecasting COVID-19 cases in Russian regions on a year in retrospect using basic statistical and machine learning methods. *Procedia Computer Science.* 2021;193:276–284. DOI: 10.1016/j.procs.2021.10.028.
3. Каширина И.Л., Ершов Д.О. Анализ, моделирование и прогнозирование COVID-19 на основе данных Воронежской области. *«Актуальные проблемы прикладной*

математики, информатики и механики»: Сборник трудов Международной конференции; 13–15 декабря 2021 г.; Воронеж. Воронеж: Издательство «Научно-исследовательские публикации»; 2022. С. 174–180.
4. Каширина И.Л., Азарнова Т.В., Бондаренко Ю.В. Анализ влияния пандемии COVID-19 на развитие человеческого капитала региона с помощью алгоритмов машинного обучения. *Моделирование, оптимизация и информационные технологии.* 2022;10(1). URL: https://moitvivt.ru/ru/journal/pdf?id=1137. DOI: 10.26102/2310-6018/2022.36.1.004 (дата обращения: 02.07.2023).
5. Ayyoubzadeh S.M., Ayyoubzadeh S.M., Zahedi H., Ahmadi M., R Niakan Kalhori S. Predicting COVID-19 incidence through analysis of Google Trends data in Iran: data mining and deep learning pilot study. *JMIR Public Health Surveill.* 2020;6(2):e18828. DOI: 10.2196/18828.
6. Venkatesh U., Aravind Gandhi P. Prediction of COVID-19 outbreaks using google trends in India: a retrospective analysis. *Healthc Inform Res*. 2020;26(3):175–184. DOI: 10.4258/hir.2020.26.3.175.
7. Pickering L., Viana J., Li X., Chhabra A., Patel D., Cohen K. Identifying factors in COVID-19 AI case predictions. In: *2020 7th International Conference on Soft Computing & Machine Intelligence (ISCMI); 2020; Stockholm, Sweden.* p. 192–196. DOI: 10.1109/ISCMI51676.2020.9311583.
8. Яковенко Н.В., Азарнова Т.В., Каширина И.Л., Бондаренко Ю.В., Щепина И.Н. *Инструментальные методы оценки человеческого капитала: Теория и прикладные аспекты.* Воронеж: Издательство «Цифровая полиграфия»; 2022. 177 с.
9. Ketu S., Mishra P.K. India perspective: CNN-LSTM hybrid deep learning model-based COVID-19 prediction and current status of medical resource availability. *Soft Comput.* 2022;26(2):645–664. DOI: 10.1007/s00500-021-06490-x.
10. Lecun Y., Bottou L., Bengio Y., Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE.* 1998;86(11):2278–2324. DOI: 10.1109/5.726791.
11. Devaraj J., Madurai Elavarasan R., Pugazhendhi R., Shafiullah G.M., Ganesan S., Jeysree A.K., Khan I.A., Hossain E. Forecasting of COVID-19 cases using deep learning models: Is it reliable and practically significant? *Results Phys.* 2021;21:103817. DOI: 10.1016/j.rinp.2021.103817.

REFERENCES

1. Braga M.D.B., Fernandes R.D.S., de Souza G.N., et al. Artificial neural networks for short-term forecasting of cases, deaths, and hospital beds occupancy in the COVID-19 pandemic at the Brazilian Amazon. *PLoS One.* 2021;16(3):e0248161. DOI: 10.1371/journal.pone.0248161.
2. Naumov A., Moloshnikov I., Serenko A., Sboev A., Rybka R. Baseline accuracies of forecasting covid-19 cases in Russian regions on a year in retrospect using basic statistical and machine learning methods. *Procedia Computer Science.* 2021;193:276–284. DOI: 10.1016/j.procs.2021.10.028.
3. Kashirina I.L., Ershov D.O. Analiz, modelirovanie i prognozirovanie COVID-19 na osnove dannyh Voronezhskoj oblasti. *Aktual'nye problemy prikladnoj matematiki, informatiki i mehaniki: Sbornik trudov Mezhdunarodnoj konferencii; 13–15 December 2021*; Voronezh. Voronezh: Nauchno-issledovatel'skie publikatsii; 2022. p. 174–180. (In Russ.).
4. Kashirina I.L., Azarnova T.V., Bondarenko Yu.V. Analysis of the COVID-19 pandemic impact on the development of human capital in the region using machine learning algorithms. *Modeling, Optimization and Information Technology.* 2022;10(1). URL:

https://moitvivt.ru/ru/journal/pdf?id=1137. DOI: 10.26102/2310-6018/2022.36.1.004 (accessed on 02.07.2023) (In Russ.).
5. Ayyoubzadeh S.M., Ayyoubzadeh S.M., Zahedi H., Ahmadi M., R Niakan Kalhori S. Predicting COVID-19 incidence through analysis of Google Trends data in Iran: data mining and deep learning pilot study. *JMIR Public Health Surveill*. 2020;6(2):e18828. DOI: 10.2196/18828.
6. Venkatesh U., Aravind Gandhi P. Prediction of COVID-19 outbreaks using google trends in India: a retrospective analysis. *Healthc Inform Res*. 2020;26(3):175–184. DOI: 10.4258/hir.2020.26.3.175.
7. Pickering L., Viana J., Li X., Chhabra A., Patel D., Cohen K. Identifying factors in COVID-19 AI case predictions. In: *2020 7th International Conference on Soft Computing & Machine Intelligence (ISCMI); 2020; Stockholm, Sweden*. p. 192–196. DOI: 10.1109/ISCMI51676.2020.9311583.
8. Jakovenko N.V., Azarnova T.V., Kashirina I.L., Bondarenko Ju.V., Shhepina I.N. *Instrumental methods of human capital assessment: Theory and practical aspects*. Voronezh, Cifrovaja poligrafija; 2022. 177 p. (In Russ.).
9. Ketu S., Mishra P.K. India perspective: CNN-LSTM hybrid deep learning model-based COVID-19 prediction and current status of medical resource availability. *Soft Comput.* 2022;26(2):645–664. DOI: 10.1007/s00500-021-06490-x.
10. Lecun Y., Bottou L., Bengio Y., Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*. 1998;86(11):2278–2324. DOI: 10.1109/5.726791.
11. Devaraj J., Madurai Elavarasan R., Pugazhendhi R., Shafiullah G.M., Ganesan S., Jeysree A.K., Khan I.A., Hossain E. Forecasting of COVID-19 cases using deep learning models: Is it reliable and practically significant? *Results Phys.* 2021;21:103817. DOI: 10.1016/j.rinp.2021.103817.

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Каширина Ирина Леонидовна, доктор технических наук, профессор, кафедра математических методов исследования операций, Воронежский государственный университет, Воронеж, Российская Федерация.
e-mail: kash.irina@mail.ru
ORCID: 0000-0002-8664-9817

Irina Leonidovna Kashirina, Doctor of Technical Sciences, Professor, Mathematical Methods of Operations Research Department, Voronezh State University, Voronezh, the Russian Federation.

Матыкина Ольга Вячеславовна, студент, Воронежский государственный университет, Воронеж, Российская Федерация.
e-mail: omatykina@mail.ru

Olga Vyacheslavovna Matykina, Undergraduate Student, Voronezh State University, Voronezh, the Russian Federation.

Статья поступила в редакцию 03.08.2023; одобрена после рецензирования 25.08.2023; принята к публикации 06.09.2023.

The article was submitted 03.08.2023; approved after reviewing 25.08.2023; accepted for publication 06.09.2023.