Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

Институт фундаментальной биологии и биотехнологии
Кафедра биофизики

УТВЕРЖДАЮ
Заведующий кафедрой биофизики
_____ В. А Кратасюк

«_____» _____2023 г.
Кафедра биофизики, Базовая кафедра
медико-биологических систем и
комплексов

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Использование искусственных нейронных сетей при прогнозировании патогенности вирусов (на примере коронавируса SARS – CoV-2)

03.04.02 Физика
03.04.02.10 Биофизика и медицинская инженерия

Научный руководитель	_____ подпись	А. Н. Шуваев инициалы, фамилия
Выпускник	___ _____ подпись	А. В. Зиненко инициалы, фамилия
Рецензент	_____ подпись	Л. В. Степанова инициалы, фамилия

Красноярск 2023

РЕФЕРАТ

Выпускная квалификационная работа по теме «Использование искусственных нейронных сетей при прогнозировании патогенности вирусов (на примере коронавируса SARS – CoV-2)» содержит 43 страницы текстового документа, 13 иллюстраций, 4 таблицы, 1 приложение, 41 использованный источник.

ВИРУЛЕНТНОСТЬ, ПАТОГЕННОСТЬ, ПРОГНОЗИРОВАНИЕ, РЕГРЕССИЯ, РЕКУРРЕНТНЫЕ НЕЙРОННЫЕ СЕТИ, АНСАМБЛЕВЫЕ АЛГОРИТМЫ, ГРАДИЕНТНЫЙ БУСТИНГ, СЛУЧАЙНЫЙ ЛЕС, СРЕДНЯЯ АБСОЛЮТНАЯ ОШИБКА В ПРОЦЕНТАХ.

Объект исследования – динамика заболеваемости и смертности от коронавируса SARS-CoV-2. Цель – применить рекуррентные нейронные сети и другие алгоритмы машинного обучения к исследованию и прогнозированию патогенности вируса SARS-CoV-2.

Задачи:

- Обзор научных работ, посвященных патогенности вирусов в целом и SARS-CoV-2 в частности.
- Обзор научных статей, посвященных прогнозированию и исследованию коронавируса SARS-CoV-2 с помощью алгоритмов машинного обучения и искусственных нейронных сетей.
- Изучение методов Data science, машинного обучения и нейронных сетей, направленных на прогнозирование и оценку влияния.
- Подготовка исходных данных по заражаемости, смертности и вакцинации по коронавирусу SARS-CoV-2 в РФ.
- Прогнозирование эволюции коронавируса SARS-CoV-2 с использованием рекурретных нейросетей.
- Оценка влияния вакцинации на смертность с использованием регрессионных нейросетей, а также ансамблевых алгоритмов.
- Интерпретация результатов.

В результате исследования рекуррентная, а именно LTSM нейросеть успешно спрогнозировала динамику заболеваемости и смертности. Ошибка составила 9,5% по новым случаям и 11,7% по новым смертям. Наличие ошибки связано с небольшим (для обучения нейросети) объемом выборки.

Регрессионные модели были построены тремя методами: рекуррентная нейросеть, градиентный бустинг и случайный лес. Наименьшую ошибку показала рекуррентная нейросеть, при этом случайный лес позволил оценить значимость факторов прироста новых случаев и вакцинации.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В настоящее время цифровизация касается абсолютно всех направлений деятельности человека. Мощное вычислительное оборудование позволяет собирать, обрабатывать и анализировать колоссальные объемы данных. В связи с этим такие понятия классической статистики как «выборка» и «выборочное наблюдение» стали не актуальными – мы имеем возможность исследовать и анализировать генеральную совокупность и делать более точные выводы. Наука о данных как дисциплина появилась еще в 1974 году, когда был введен термин Data Science, но современный вид приняла в начале двухтысячных с появлением вычислительных мощностей, способных обрабатывать большие данные. На данный момент помимо вычислительных мощностей мы имеем доступную информацию, и можем получать необходимые данные из открытых источников.

С понятием Data science тесно связано понятие машинного обучения. Машинное обучение – это поиск закономерностей в массиве представленной информации и выбор наилучшего решения без участия человека. Для реализации машинного обучения используются в том числе искусственные нейросети.

Data science широко применяется не только в исследовательских, но и в прикладных целях. Особенно активно используют анализ данных в бизнесе, маркетинге и финансах. Также, с развитием современных технологий в медицине, все больше медицинских проблем решается Data science. Это, к примеру, распознавание изображений, построение моделей распространения эпидемии, предсказание заболеваний, генетические исследования.

В магистерской диссертации исследовалась патогенность и вирулентность вируса SARS-CoV-2 и прогнозировалась его эволюция с использованием алгоритмов машинного обучения. Патогенность и вирулентность – это близкие понятия, но они отличаются тем, что патогенность – это способность вируса наносить вред организму, а вирулентность – это способность вируса заражать, в том числе и бессимптомно.

Основной частью работы выступило исследование влияния на смертность от коронавирусной инфекции вакцинации в Российской Федерации. Для сравнения также исследуется значимость фактора вакцинации в странах, в которых порог коллективного иммунитета был достигнут. Помимо вышеуказанного, в работе был произведен прогноз заболеваемости и смертности от нового коронавируса в РФ с использованием рекуррентной нейронной сети и проверена его результативность. Также был вылнен прогноз регриссионным методом с использованием рекуррентной нейронной сети и ансамблевых алгоритмов.

Цель данной работы – применить рекуррентные нейронные сети и другие алгоритмы машинного обучения к исследованию и прогнозированию

патогенности вируса SARS-CoV-2. Для достижения поставленной цели решены следующие **задачи**.

- Обзор научных работ, посвященных патогенности вирусов в целом и SARS-CoV-2 в частности.
- Обзор научных статей, посвященных прогнозированию и исследованию коронавируса SARS-CoV-2 с помощью алгоритмов машинного обучения и искусственных нейронных сетей.
- Изучение методов Data science, машинного обучения и нейронных сетей, направленных на прогнозирование и оценку влияния.
- Подготовка исходных данных по заражаемости, смертности и вакцинации по коронавирусу SARS-CoV-2 в РФ.
- Прогнозирование эволюции коронавируса SARS-CoV-2 с использованием рекурретных нейросетей.
- Оценка влияния вакцинации на смертность с использованием регрессионных нейросетей, а также ансамблевых алгоритмов.
- Интерпретация результатов.

Объектом исследования является динамика заболеваемости и смертности от коронавируса SARS-CoV-2, предметом – применение нейронных сетей для прогнозирования патогенности вируса. В работе использованы методы машинного обучения, искусственные нейросети, статистические методы прогнозирования.

Научная новизна исследования состоит в том, что методы машинного обучения были применены к временным рядам распространения и смертности, впервые было оценено влияние вакцинации на смертность от вируса с использованием ансамблевых методов машинного обучения. Инструментами исследования выступают современные IT технологии, в частности искусственные нейронные сети. Направленность работы – теоретическая, но полученные результаты можно использовать в прикладных исследованиях.

Источниками информации выступили учебные пособия по вирусологии и анализу данных, статьи в научных журналах, а также интернет – ресурсы.

ЗАКЛЮЧЕНИЕ

Исследование выполнялось на примере Российской Федерации за период с марта 2020 по сентябрь 2022 г. Основные задачи исследования были следующие.

- Осуществить интервальный прогноз заболеваемости и смертности от Covid-19 с использованием LTSM - нейросети и проверить точность данного прогноза
- Построить нейросетевую регрессионную модель зависимости смертности от вакцинации и заражаемости различными методами и сделать выводы по результатам модели.
- Сравнить результаты оценки влияния вакцинации в РФ со странами, в которых был достигнут порог коллективного иммунитета.

Все задачи были успешно выполнены. В результате исследования рекуррентная, а именно LTSM нейросеть успешно спрогнозировала динамику заболеваемости и смертности. Ошибка составила 9,5% по новым случаям и 11,7% по новым смертям. Наличие ошибки связано с небольшим (для обучения нейросети) объемом выборки.

Регрессионные модели были построены тремя методами: рекуррентная нейросеть, градиентный бустинг и случайный лес. Наименьшую ошибку показала рекуррентная нейросеть, при этом случайный лес позволил оценить значимость факторов прироста новых случаев и вакцинации.

Выпускная квалификационная работа имеет как теоретическую, так и прикладную направленность. Теоретическая значимость работы состоит в применении алгоритмов машинного обучения к прогнозированию эволюции пандемии и к оценке влияния факторов на смертность от пандемии. Практическая значимость заключается в том, что результаты прогноза можно использовать для разработки профилактических мероприятий. Также интересны с практической точки зрения результаты регрессии, которые показали, что вакцинация весьма существенно влияет на смертность от пандемии, как в Российской Федерации, так и в странах, в которых население дисциплинированно прививалось.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Вирусология: учебник / А. В. Пиневич, А. К. Сироткин, О. В. Гаврилова, А. А. Потехин; под ред. А. В. Пиневича – Санкт Петербург: Изд-во С.-Петерб. ун-та, 2020. – 442 с.
2. Гудфелло Я. Глубокое обучение / Я. Гудфеллоу, И. Бенджио, А. Курвилль; пер. с анг. А. А. Слинкина. – Москва: ДМК, 2018. – 674 с.
3. Данилов А. В. Технология электронного бенчмаркинга медицинских организаций региона/ А. В. Данилов // Цифровое здравоохранение. Труды XX международного конгресса «Информационные технологии в медицине». – Москва: Консэф, 2019 – С. 28 – 30.
4. Ильин, В.А. Магистральные направления физики XXI века: Физика технологий будущего для будущих физиков и инженеров: Современная макрофизика: Низкие температуры. Сверхпроводимость. Сверхтекучесть. Лазеры. Фуллерены, нанотрубки, графен. Информационные технологии / В. А. Ильин, В. В. Кудрявцев. –Москва: Ленанд, 2018. – 448 с.
5. Кабанихин С. И., Криворотько О.И. Математическое моделирование эпидемии Уханьского коронавируса COVID-19 и обратные задачи/ С.И. Кабанихин, О. И Криворотько // Журнал вычислительной математики и математической физики. – 2020. – т.60, № 11. – С. 1950 – 1961.
6. Карпова О.В. Вирусология. Конспект лекций / О.В. Карпова. – Москва: Фонд «Вольное дело», 2020. – 162 с.
7. Куркина Е. С., Кольцова Е. М. Математическое моделирование и прогнозирование распространения эпидемии коронавируса COVID-19/ Е. С. Куркина, Е. М. Кольцова // Проектирование будущего. Проблемы цифровой реальности: труды 4-й Международной конференции (4-5 февраля 2021 г., Москва). — Москва: ИПМ им. М.В.Келдыша, 2021. — С. 178-192.
8. Литусов Н.В., Устюжанин А.В. Структура и репродукция вирусов. Иллюстрированное учебное пособие / Н.В. Литусов, А.В. Устюжанин. – Екатеринбург: Изд-во УГМА, 2012. – 209 с.
9. Маракулин И.В. Медицинская микробиология. Курс лекций: учебное пособие / И.В. Маракулин. – Киров: ФГБОУ ВПО «ВятГУ», 2011. – 119 с.
10. Медицинская вирусология: учебное пособие / И.И. Генералов, Н.В. Железняк, В.К. Окулич и др.; под ред. И.И. Генералова. – Витебск: ВГМУ, 2017. – 307 с.
11. Невзоров В. П. Комплексное использование алгоритмов нейронных сетей для оценки эффективности их работы в меняющихся условиях принятия врачебных решений / В.П. Невзоров, Т.М. Буланова // Цифровое здравоохранение. Труды XX международного конгресса «Информационные технологии в медицине». – Москва: Консэф, 2019 – С. 10 – 13.

12. Новикова Н.А. Молекулярные аспекты взаимодействия вирусов с клеткой: Учебное пособие/ Н. А. Новикова. – Нижний Новгород: Изд-во ННГУ им. Н.И. Лобачевского, 2015. – 87 с.
13. Рындина С.В. Бизнес-аналитика на основе больших данных: обучение без учителя на языках Python и R : учеб.-метод. пособие / С. В. Рындина. – Пенза: Изд-во ПГУ, 2020. – 76 с.
14. Супотницкий М.В. Новый коронавирус SARSCoV-2 в аспекте глобальной эпидемиологии коронавирусных инфекций / М.В. Супотницкий // Вестник войск РХБ защиты. – 2020. – Т. 4, № 1. – С. 32–65.
15. Суханова Н.В. Разработка нейросетевой модели для мониторинга заболеваемости и прогнозирования эффективности противоэпидемических мер / Н.В. Суханова // Вестник Брянского технического университета. – 2020. – № 10 (95). – С. 42 –50.
16. Федосов В. В., Федосова А. В., Buitrago O. Стохастическая оценка развития эпидемии в локальных скоплениях населения/ В. В. Федосов, А. В. Федосова, O. Buitrago // Информационные технологии. – 2021. – т. 27, № 8. – С. 435 – 444.
17. Хайтович А.Б. Коронавирусы (структура генома, репликация) / А.Б. Хайтович // Крымский журнал экспериментальной клинической медицины. – 2020– № 10 – С. 79–95.
18. Шоле Ф. Глубокое обучение на Python / Франсуа Шоле. – Санкт Петербург: Питер, 2018. – 400 с.
19. A COVID-19 Pandemic Artificial Intelligence–Based System With Deep Learning Forecasting and Automatic Statistical Data Acquisition: Development and Implementation Study // Ch.-Sh. Yu et al. – Journal of Medical Internet Research. – 2021. – vol. 25 № 5.
20. A Deep Learning Method to Forecast COVID-19 Outbreak / S. Dash et al. // New Generation Computing. – 2021. – vol. 39– P. 515–539.
21. Al-Najjar D. Evaluation of the prediction of COVID-19 recovered and unrecovered cases using symptoms and patient's meta data based on support vector machine, neural network, CHAID and QUEST Models//D. Al-Najjar1, H. Al-Najjar, N. Al-Nousan.// European Review for Medical and Pharmacological Sciences. – 2021. – vol 25. – P. 5556 – 5560.
22. Ben Yahia N. Integrating Models and Fusing Data in a Deep Ensemble Learning Method for Predicting Epidemic Diseases Outbreak / N. Ben Yahia, M. Kandara, N. Ben Saoud // Big Data Research. – 2022. – vol. 27.
23. Big data in healthcare: management, analysis and future prospects/ S. Dash, S. K. Shakiavar et al. // Journal of Big Data. – 2019. – № 6. – URL: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0217-0.
24. Brockmann D. The Hidden Geometry of Complex, Network-Driven Contagion Phenomena / D.Brockmann, D. Helbing // SCIENCE. – 2013. – Vol 342, Issue 6164. – P. 1337-1342.

25. Chimmula V.K. Time series forecasting of COVID-19 transmission in Canada using LSTM Networks/ V.K. Chimmula, L. Zhang // Chaos, Solitons And Fractals. – 2020. –vol 35.
26. Convolutional neural networks and temporal CNNs for COVID-19 forecasting in Franc e/ L. Mohimont et al. // Applied Intelligence. – 2021. – vol. 51 – P. 8784–8809 .
27. COVID-19 in Saudi Arabia / A.M. Ajbar, M. Ali, A. Ajbar // The Journal of Infection in Developing countries. – 2021– 15(7). – P. 918-924.
28. Dadyan E. Neural Networks and Forecasting COVID-19 / E. Dadyan, P. Avetsyan // Optical Memory and Neural Networks. – 2021. – vol. 30. – P. 225–235.
29. Deep learning-based forecasting model for COVID-19 outbreak in Saudi Arabia./ A.H. Elsheikh et al. // Process Safety and Environmental Protection.– 2021. – volume 149– P. 223-233.
30. Enhanced bat algorithm for COVID-19 short-term forecasting using optimized LSTM / H.T. Rauf et al. // Soft Computing. – 2021. – vol. 25. – P. 12989–12999
31. Ferguson N. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand / N. Ferguson et al. // Imperial College COVID-19 Response Team. – March 2020. – URL: https://www.researchgate.net/publication/342182508
32. Improving prediction of COVID-19 evolution by fusing epidemiological and mobility data /S. Garcia-Gremades et al. // Scientific Reports. – 2021. – № 11.
33. Interventions to control nosocomial transmission of SARS-CoV-2: a modelling study/ T.M. Phan, H. Tahir et al. // BMC Medicine. – 2021. – vol.19, iss. 211. – URL: https://doi.org/10.1186/s12916-021-02060-y.
34. Lounis M.Predictive models for COVID-19 cases, deaths and recoveries in Algeria / M. Lounis, O. Torrealba-Rodriguez, R. A. Conde-Gutiérrez // Results In Physics.– 2021 – vol. 30.
35. Melin P. Spatial and temporal spread of the COVID-19 pandemic using self organizing neural networks and a fuzzy fractal approach / P. Mellin, O. Castillo // Sustainability (Switzerland). – vol. 13, iss. 50.
36. Niazkar H.R. Application of artificial neural networks to predict the COVID-19 outbreak/ H.R. Niazkar, M. Niazkar // Global Health Research and Policy. – 2020. –vol 5(50). – URL: https://rdcu.be/cAE8Y
37. Performance Evaluation of Soft Computing Approaches for Forecasting COVID-19 Pandemic Cases/ M. Soahib et al. // SN Computer Science. – 2021. – vol 2, iss. 372.
38. Recurrent Neural Network and Reinforcement Learning Model for COVID-19 Prediction / L.R. Kumar et al. // Front. Public Health. – 2021.
39. Shetty R. Forecasting of COVID 19 Cases in Karnataka State using Artificial Neural Network (ANN)/ R. Shetty, P. Srinivasa //Journal of The Institution of Engineers (India): Series B. – 2021. – vol. 1. – P. 1201 – 1211.

40. Spatio-temporal prediction of the COVID-19 pandemic in US counties: modeling with a deep LSTM neural network / B. Nikparvar et al. // Scientific Reports. – 2021 – № 11.
41. Younis M. Ch. Evaluation of deep learning approaches for identification of different corona-virus species and time series prediction / M.Ch. Younis // Computerized Medical Imaging and Graphics. – 2021. –vol. 90.

ПРИЛОЖЕНИЕ А. Классификация вирусов Д. Балтимора

Группа		Тип нуклеиновой кислоты	Тип симметрии	Наличие суперкапсида	Основные семейства
I		двунитчатая ДНК линейная	Кубический	Нет	Аденовирусы
		двунитчатая ДНК линейная	Комплексный	Да	Герпесвирусы
		двунитчатая ДНК линейная	Кубический	Да	Поксвирусы
		двунитчатая ДНК кольцевая	Кубический	Нет	Полиомавирусы, папилломавирусы
II		однонитчатая ДНК линейная	Кубический	Нет	Парвовирусы
		однонитчатая ДНК кольцевая	Кубический	Нет	Анелловирусы
III		двунитчатая РНК сегментированная	Кубический	Нет	Пикобирнвирусы, реовирусы
IV		однонитчатая РНК(+)	Кубический		Пикорнавирусы, астровирусы, калицивирусы, гепевирусы
		однонитчатая РНК(+)	Кубический	Да	Тогавирусы, флавивирусы
		однонитчатая РНК(+)	Спиральный	Да	Коронавирусы

Группа		Тип нуклеиновой кислоты	Тип симметрии	Наличие суперкапсида	Основные семейства
V		однонитчатая РНК(-)	Спиральный	Да	Парамиксовирусы, филовирусы, борнавирусы, рабдовирусы
		однонитчатая РНК (-) сегментированная	Спиральный	Да	ортомиксовирусы
		однонитчатая РНК(-) сегментированная	Спиральный или кубический	Да	буньявирусы
		однонитчатая РНК(-) сегментированная амбиполярная	Не определен	Да	аренавирусы
VI		однонитчатая РНК(+) 2 копии обратная транскрипция	Капсид конусовидной формы	Да	ретровирусы
VII		двунитчатая ДНК кольцевая обратная транскрипция	Кубический	Да	

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

Институт фундаментальной биологии и биотехнологии
Кафедра биофизики

УТВЕРЖДАЮ
Заведующий кафедрой Биофизики
В.А. Кратасюк _____
«15» 06 2023 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Использование искусственных нейронных сетей при прогнозировании патогенности вирусов (на примере коронавируса SARS – CoV-2)

03.04.02 Физика
03.04.02.10 Биофизика и медицинская инженерия

Научный руководитель	подпись	15.06.23.	А. Н. Шуваев инициалы, фамилия
Выпускник	подпись	15.06.23.	А. В. Зиненко инициалы, фамилия
Рецензент	подпись	15.06.23.	Л. В. Степанова инициалы, фамилия

Красноярск 2023