



**«Московский государственный технический университет
имени Н.Э. Баумана»**

(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ _____ Информатика и системы управления _____

КАФЕДРА _____ Системы обработки информации и управления _____

РАСЧЁТНО - ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

к курсовой работе по дисциплине «Обработка и анализ данных» на тему:

_____ Анализ представления медицинских данных пациентов больных
COVID-19 для обучения искусственного интеллекта

Студент группы _____ ИУ5-34М _____ / _____ Ишков Д.О. _____
(Подпись, дата) (И.О.Фамилия)

Руководитель курсового проекта _____ / _____ Гапанюк Ю.Е. _____
(Подпись, дата) (И.О.Фамилия)

Москва, 2021

Оглавление

Глоссарий	3
Введение	3
Постановка задачи	5
Описание форматов представления	6
DICOM	6
Описание набора данных для обучения	8
Датасет	8
Определение критериев качества рентгенографических изображений	9
Выбор методов и средств искусственного интеллекта	11
Результаты	13
Оценка полученных результатов и дисперсионный анализ	16
Заключение	22
Список использованной литературы	23

Глоссарий

1. ИИ – искусственный интеллект
2. SARS-Cov-2 - Severe acute respiratory syndrom -related coronavirus 2 или тяжёлым острым респираторным синдромом вызванный коронавирусом
3. COVID-19 – инфекционное заболевание, вызываемое коронавирусом
4. КТ – компьютерная томография
5. ПЦР – полимеразная цепная реакция
6. DICOM – Digital Imaging and Communications in Medicine, медицинский отраслевой стандарт создания, хранения, передачи и визуализации цифровых медицинских изображений и документов обследованных пациентов
7. HU – единицы шкалы Хаунсфилда
8. ГБ – гигабайт или 10^9 байт
9. ROI – Region of interest, область интереса

Введение

Коронавирус COVID-19 – это инфекционное заболевание, вызываемое вирусом SARS-CoV-2. Как следует из названия, вирус характеризуется SARS (Severe acute respiratory syndrome) или тяжёлым острым респираторным синдромом, в пять раз смертоноснее, нежели вирус гриппа, COVID-19 поражает дыхательную систему, вызывая кашель, одышку и смерть в тяжелых случаях. Первый случай заболевания был зарегистрированный 31 декабря 2019 года, тогда как уже 11 маёрта 2020 года Всемирная

Организация Здравоохранения объявила о пандемии новой коронавирусной инфекции. За последние годы было обнаружено множество более заразных и летальных штампов, среди которых следует отметить штамм B.1.617.2 (Дельта), ставшим наиболее распространённым во второй половине 2021 года, а также штамм B.1.1.529 (Омикрон), который был обнаружен в ноябре 2021 года. Для диагностики SARS-CoV-2 используются тесты полимеразной цепной реакции (ПЦР). Доля ложноотрицательных результатов (False Negative Rate) этих тестов высока, поэтому для точного диагноза необходимы дополнительные методы. Кроме того, может пройти несколько часов, а иногда и дней, прежде чем будут получены результаты молекулярного анализа.

Одним из эффективных способов диагностики формы и тяжести болезни являются рентгенографические снимки и компьютерная томография (КТ) грудной клетки. КТ обеспечивает детальное обследование грудной клетки для диагностики SARS-CoV-2 и состоит из сотен двухмерных срезов. Хотя и существуют руководящие принципы, помогающие рентгенологам и пульмонологам отличать коронавирус от других типов инфекции, их оценки могут различаться.

Для повышения эффективности диагностирования инфекции с высокой точностью необходим метод, основанный на искусственном интеллекте (ИИ). В условиях ограниченных ресурсов, медицинского оборудования и специалистов предлагается применить свёрточные нейронные сети, которые зарекомендовали себя в ряде задач компьютерного зрения, в том числе и классификации заболеваний по снимкам [2-5]. Для сравнения также будет проведено обучение с помощью ансамбля деревьев решений, благодаря которому становится возможным локализовать участки поражений лёгких.

Постановка задачи

Предметом анализа в данной работе являются рентгеновские цифровые снимки и двухмерные срезы компьютерной томографии лёгких в формате DICOM больных атипичной формой пневмонии, выраженной COVID-19.

Цель работы заключается в получение ответа на вопрос, какого качества изображения компьютерной томографии лёгких достаточно для эффективного по заданной точности и времени распознавания COVID-19 с помощью моделей-классификаторов, основанных на свёрточных нейронных сетях.

Алгоритм анализа:

Шаг 1. Формирование математической модели качества фотографий снимков грудной клетки как интегрального критерия по их параметрам

Шаг 2. Выбор альтернативных вариантов математических моделей для сравнения

Шаг 3. Планирование набора исследуемых критериев изображения для анализа

Шаг 4. Определение и обоснование числа уровней для каждого из критериев

Шаг 5. Формирование плана проведения эксперимента и обучения моделей

Шаг 6. Выбор архитектуры свёрточной нейросети и её гиперпараметров

Шаг 7. Описание обучения модели

Шаг 8. Экспериментальная оценка качества обученными моделями

Шаг 9. Проведение регрессионного и дисперсионного анализа

Шаг 10. Выводы, которые можно извлечь из полученных результатов

Описание форматов представления

DICOM

DICOM (Digital Imaging and Communications in Medicine) – медицинский отраслевой стандарт создания, хранения, передачи и визуализации цифровых медицинских изображений и документов обследованных пациентов [6]. Используется согласно стандарту ISO 12052:2017 во всех медицинских учреждениях. Особенности формата – черно-белые изображения с метаданными, где один пиксель имеет 16 бит, из которых 4 зарезервированы, или 4096 цветов.

Выбор 12 бит для целочисленного беззнакового хранения пиксельных данных DICOM файла обусловлен тем фактом, что медицинское оборудование дискретизирует измеряемые аналоговые величины по шкале Хаунсфилда (HU) – шкала линейного ослабления излучения по отношению к дистиллированной воде:

$$HU(\mu_x) = \frac{\mu_x - \mu_{water}}{\mu_{water} - \mu_{air}} \times 1000, (1)$$

где μ_{water} и μ_{air} – линейные коэффициенты ослабления для воды и воздуха при стандартных условиях. Шкала была предложена сэром Годфри Ньюболдом Хаунсфилдом, одним из главных инженеров и разработчиков аксиальной компьютерной томографии. Для воздуха значение по шкале HU составляет минус 1000 единиц, для воды 0 единиц, для костей 400 единиц и выше, вплоть до 3000 единиц HU (см. рисунок 1). Поэтому для грамотного представления интенсивности в цифровом формате по этой шкале необходимо как минимум 12 бит. При этом стоит иметь в виду, что,

поскольку данных хранятся в формате без знака, то 0 у пикселя в DICOM файле соответствует минус 1000 единиц HU.

С учётом того, что большинство рабочих мониторов имеют цветность 8 или 10 бит на цветовой канал, то использовать 16 бит в целочисленном беззнаковом формате на черно-белый пиксель для хранения может оказаться избыточным, поскольку на экране оператор всё равно не увидит различия между 16 и 10 битным черно-белым изображением. Кроме того, программные средства для обработки и редактирования цифровых изображений работают с форматом 8 бит на один цветовой канал.

Для обучения моделей компьютерного зрения предлагается использовать всю и частичную цветность изображений. В компьютерах целочисленный формат без знака будет преобразован в числовой формат с плавающей точкой 16 бит, 32 или 64 бит (стандарт IEEE 754), т.е. без ошибок в преобразовании для 32 и 64 бит, который и используется по умолчанию для обучения моделей.



Рисунок 1. Шкала линейного ослабления рентгеновского излучения Хаунсфилда

Описание набора данных для обучения

Датасет

Как ведущая организация здравоохранения в своей области, миссия Общества информатики визуализации в медицине (SIIM) заключается в продвижении информатики медицинской визуализации посредством образования, исследований и инноваций. Для создания набора данных SIIM сотрудничает с Фондом содействия здравоохранению и биомедицинским исследованиям региона Валенсия (FISABIO), Банком данных медицинской визуализации региона Валенсия (BIMCV) и Радиологическим обществом Северной Америки (RSNA).

В качестве набора данных для обучения предлагается использовать SIIM-FISABIO-RSNA COVID-19. В этом датасете ставится задача по выявлению и локализации аномалий COVID-19 на КТ грудной клетки. В частности, стоит задача классификации КТ как отрицательные на пневмонию или типичные, неопределенные или нетипичные для COVID-19.

В набор данных входят 6334 файла для обучения с КТ лёгких в формате DICOM разрешением не менее 2000 пикселей в ширину и не менее 2700 пикселей в высоту. Общий размер набора данных составляет 120 ГБ, на рисунке 2 приведена гистограмма веса файлов в МБ на цифровом носителе информации.

В формате DICOM так же содержится информации о поле пациента, которому принадлежит снимок КТ. Эту информацию предлагается использовать для стратифицированного разделения на обучающую и валидационную выборки.

Определение критериев качества рентгенографических изображений

В качестве факторов, влияющих на качество изображения как такового, можно выделить множество вариантов. Одним таким набором является глубина цвета, разрешение изображения и погрешностей измерений технического оборудования. Обозначив двоичный логарифм размер изображения за R , двоичный логарифм числа оттенков за D , ошибки во время измерений за E , можно сформулировать интегральный критерий качества как кортеж $\langle R, D, E \rangle$. Очевидно, что на качество влияет не столько размер стороны изображения, сколько его площадь, а значит, качество зависит от квадрата R . Также очевидно, что погрешности измерений негативно влияют на качество изображения. Нормализуем значения факторов в диапазон от 0 до 1 путём нормализации по минимальному и максимальному значениям. Предложим интегральный критерий качества в виде аддитивной суммы в качестве начального приближения математической модели:

$$Q(R_{norm}, D_{norm}, E_{norm}) = R_{norm}^2 + D_{norm} - E_{norm}$$

Где R_{norm} – нормализованный размер одной из сторон радиографического изображения, D_{norm} – нормализованная цветность изображения, E_{norm} – нормализованные ошибки во время измерений.

Существует много способов классифицировать ошибки рентгенографических и КТ снимков. В качестве одной из классификаций, предложенной в работе [9], выделяются следующие группы ошибок, влияющие на E_{norm} :

1. Артефакты – ювелирные изделия, пояс или другие рентгеноконтрастные предметы, оставшиеся на пути луча, а также царапины или грязь на пластине или в кассете. В эту группу также

включены двойные экспозиции одних и тех же или разных пациентов на одной пластине.

2. Неправильное расположение – смещенные изображения, неправильная коллимация или неправильное расположение кассеты, что приводит к вырезанию изображения.
3. Движение – движение пациента или движение изображаемой области (неправильная задержка дыхания).
4. Плохая техника – недостаточная экспозиция снимка. Неправильная экспозиция может быть вызвана неправильным выбором параметров экспозиции или неправильным расположением пациента по отношению к автоматическим экспозиционным камерам.
5. Обработка – поврежденные изображения из-за сбоев в процессе оцифровки или архивирования (или во время передачи по сети).

Согласно исследованию 2020 года [10] из 90 тысяч изображений группой экспертов по выше приведённым категориям было отбраковано порядка 9% снимков. Что даёт нам математическое ожидание теоретически возможного качества изображения в 91%. Это означает, что мы можем выразить E_{norm} через постоянную систематическую ошибку, заменив его в формуле на свободный член β . Очевидно, что для свёрточной нейросети с фиксированным числом степеней свободы при бесконечном росте разрешения или глубины цвета не будет наблюдаться улучшения качества. Поэтому интегральный критерий качества изображения зависит нелинейно от локальных показателей и предлагается добавить демпфирующий коэффициент, аналогично тому, как он используется при обработке текстов в векторном представлении BM25 [11]. Так же стоит учесть, что вклад разрешения и цветности может быть неравным, поэтому следует ввести весовые коэффициенты. С учётом всего выше перечисленного получим окончательный вид качества изображения как:

$$Q(R_{norm}, D_{norm}) \cong \alpha \left(\frac{R_{norm}}{a1 + R_{norm}} \right)^2 + \gamma \frac{D_{norm}}{a2 + D_{norm}} + \beta, (2)$$

где α – весовой коэффициент, характеризующий вклад разрешения в качество, γ – весовой коэффициент, характеризующий вклад числа оттенков в качество, $a1$ – демпфирующий коэффициент для разрешения, $a2$ – демпфирующий коэффициент для числа оттенков, β – свободное слагаемое, выражающее систематическую ошибку.

Выбор методов и средств искусственного интеллекта

Для оценки качества прогнозирующей способности предлагается использовать комплексную метрику бинарной классификации третьего уровня [1]. Для её расчета необходимо рассчитать показатели первого и второго уровней. В качестве показателей первого уровня используются значения из матрицы ошибок для различных порогов. На их основе затем получают показатели второго уровня, а именно чувствительность (Sensitivity) и специфичность (Specificity):

$$Sensitivity(thr) = \frac{TP(thr)}{TP(thr) + FN(thr)},$$

$$Specificity(thr) = \frac{TN(thr)}{TN(thr) + FP(thr)},$$

где TP – число раз, когда модель верно отнесла объект к положительному классу, FN – ошибка второго рода, FP – ошибка первого рода, TN – число раз, когда модель верно отнесла объект к отрицательному классу. Данные показатели зависят от выбранного порогового значения thr и модели-классификатора.

В качестве комплексной метрики третьего уровня считается ROC-AUC, площадь под ROC-кривой – зависимости чувствительности от единицы минус специфичности на разных порогах. При этом число точек кривой не

превосходит число объектов. Значение площади под ней (AUC, area under curve), принимает значения в диапазоне от 0,5 до 1, поскольку для алгоритма случайного угадывания зависимость чувствительности от специфичности линейная.

В качестве модели свёрточной нейронной сети предлагается использовать архитектуру MobileNetV2 [12] с гиперпараметром alpha равным 1 со случайной инициализацией параметров. В данной конфигурации модель имеет порядка двух с половиной миллионов параметров для обучения. В качестве функции потерь используется бинарная перекрёстная энтропия, тем самым решается задача бинарной классификации по наличию или отсутствию пневмонии. Т.е. случаи атипичной, типичной и промежуточной пневмонии представляются одним классом «наличие пневмонии».

Модель обучается с помощью градиентного спуска, 48 эпох с размером минибатча 16 входных изображений с метками классов. Шаг обучения 10^{-3} , после 40 эпох уменьшается на порядок. В качестве оптимизатора используется Adam [13]. Входные изображения минибатча во время обучения аугментируются путем различных поворотов, смещений, отражения относительно вертикальной оси, аффинных преобразований, применения контрастности и яркости, а также нормального шума.

Очевидно, что исследуемая зависимость качества от разрешения и цветности нелинейная, поэтому предлагается использовать поиск полным перебором. Для оценки качества модели в одной конфигурации используется скользящий контроль: на 80% данных модель будет обучаться, а на остальных 20% будет измеряться её качество. Затем выбирается другое разделение на подвыборки. В итоге качество модели в одной конфигурации оценивается на 5 разделениях, а затем рассчитанные 5 значений качества ROC-AUC усредняются. Выборки же разделяются путём стратификации по целевому признаку, а также по полу пациента.

Для разрешений было решено выбрать 10 уровней: 1x1, 2x2, 4x4, 8x8, 16x16, 32x32, 64x64, 128x128, 256x256 и 512x512 пикселей. Поскольку нас интересует влияние цветности при низких её значениях, было решено выбрать 7 уровней: 2, 4, 8, 16, 64, 256 и 4096, соответственно. Таким образом, для ответа на вопрос, влияют ли факторы разрешения и градаций серого на качество распознавания свёрточной нейронной сети, необходимо провести 350 экспериментов полным перебором.

Для полной цветности на разных разрешениях были обучены более простые модели нелинейной аппроксимации: ансамбли деревьев решений для получения базовой оценки качества. Ансамбль состоит из 100 деревьев с максимальной глубиной 20. Глубина дерева была выбрана оптимально для всех разрешений путём предварительного перебора кандидатов. Количество деревьев было выбрано как оптимальное по времени обучения и достигаемому качеству. Было проведено дополнительное исследование деревьями решений на то, как влияют различные пороги изображения на окончательное качество картинки.

Результаты

На рисунке 2 представлена экспериментальная оценка поверхности отклика качества изображения от разрешения и числа оттенков серого.

Как видно из графика цветность изображения незначительно влияет на его качество, необходимое для корректной классификации пневмонии в лёгких. Лучшего качества в 87% распознавания удалось достигнуть при конфигурации разрешения 256x256 пикселей и числа градаций серого, равного 256. Для 2 оттенков – бинарного изображения, при отсечении по пороговому значению старшего бита (для 12 битных оригинальных изображений соответственно значение 12 бита), получилось достичь качества 77%.

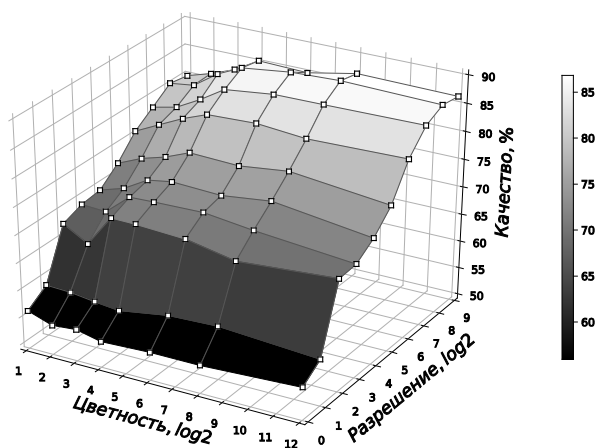


Рисунок 2. Полученные результаты обучения свёрточной нейросети MobileNetV2

Путём обучения ансамблей деревьев решений было изучено влияние выбранного порога для бинарных изображений на различных разрешениях.

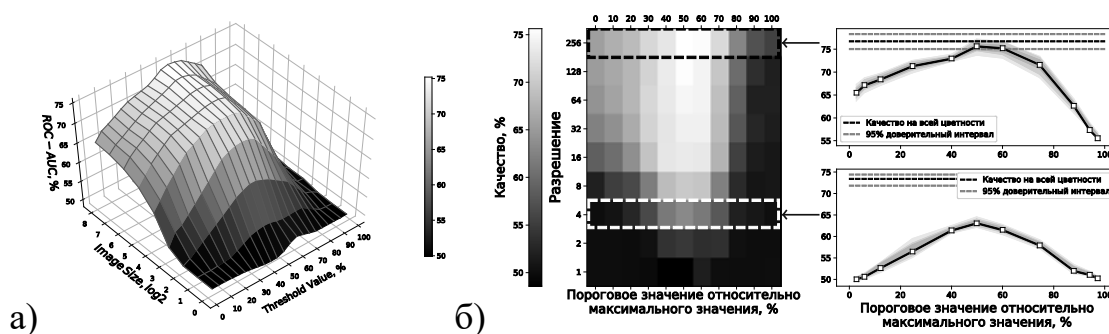


Рисунок 3. Полученные результаты обучения ансамбля деревьев решений на бинарных данных для разных разрешений и порогов: а) трёхмерное представление; б) двухмерное представление с помощью горячей карты и срезов на отдельных разрешениях;

Как видно из результатов, бинаризация по старшему биту даёт наилучшие результаты для стандартной бинаризации. Данный подход не позволяет сравниться с качеством ансамбля, обученного на всей цветности (пунктирная линия). Полученные результаты в 76% сравнимы с результатами, полученными при обучении свёрточной нейросети на бинарных изображениях.

Также был поставлен альтернативный эксперимент по адаптивной бинаризации изображений при помощи электронного цифрового фильтра Гаусса, чьей импульсной переходной функцией является функция Гаусса. Фильтр Гаусса спроектирован таким образом, чтобы не иметь перерегулирования в переходной функции и максимизировать постоянную времени. Фильтр Гаусса (Gaussian filter) обычно используется в цифровом виде для обработки двумерных сигналов (изображений) с целью снижения уровня шума (см. рисунок 4). Результаты для некоторых разрешений представлены на рисунке 5.

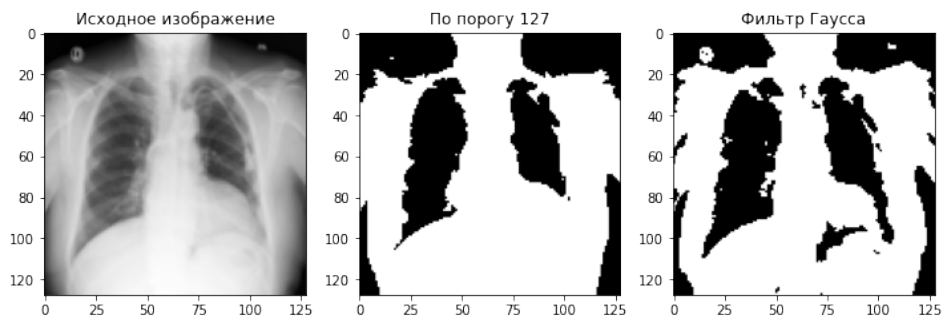


Рисунок 4. Различия между обычной и адаптивной бинаризацией

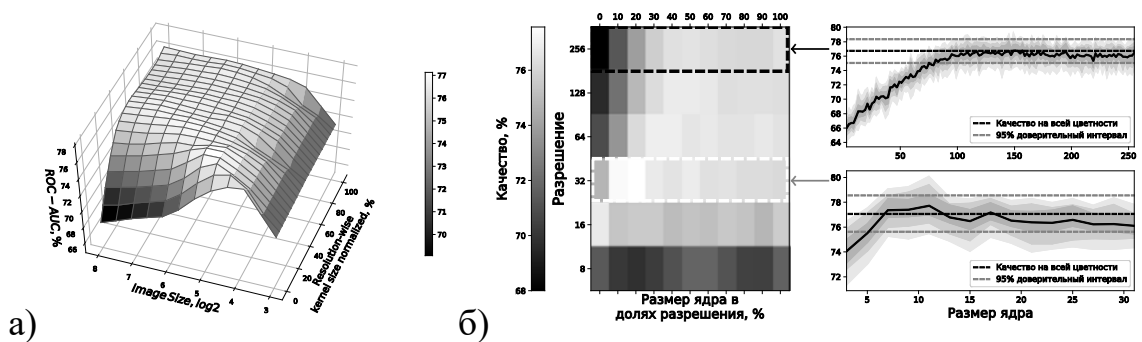


Рисунок 5. Полученные результаты обучения ансамбля деревьев решений на бинарных данных для разных разрешений и ядер фильтра Гаусса:

- а) трёхмерное представления; б) двухмерное представление с помощью горячей карты и срезов на отдельных разрешениях;

Оценка полученных результатов и дисперсионный анализ

Результаты, отображенные на рисунке 2, были обработаны и представлены в двухмерных срезах на рисунке 6. По рисунку видно, что модель случайного леса при размере изображения до 64 пикселей превосходит по качеству сверточную модель. При дальнейшем увеличении разрешения качество случайного леса не улучшается, в отличие от качества MobileNetV2. Качество MobileNetV2 перестает увеличиваться при размере разрешения 512 пикселей. Количество градаций пикселя незначительно влияет на качество изображения. Статистически значимые различия в средних получены лишь для 2 и 4 оттенков.

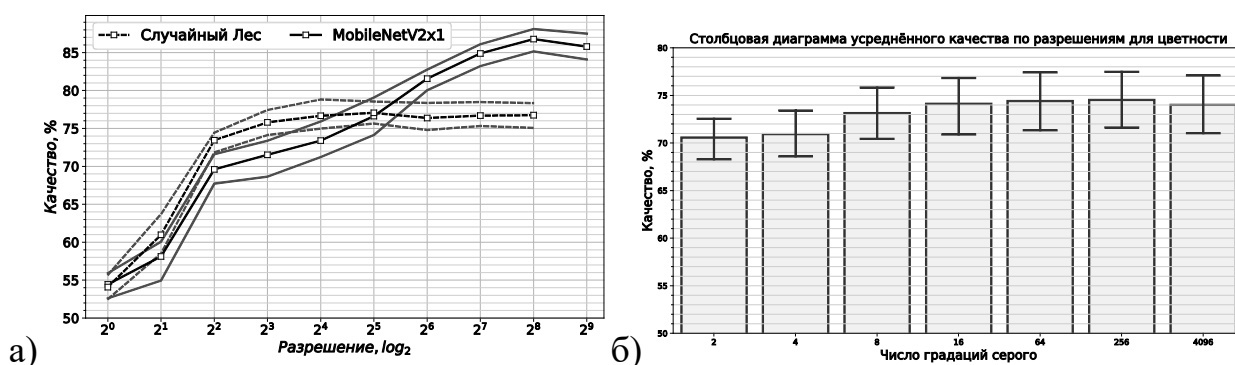


Рисунок 6. Обработка экспериментальных зависимостей: а) зависимость качества от разрешения для моделей, обученных на 256 оттенках, с 95% доверительными интервалами матожидания; б) столбчатая диаграмма усреднённого качества MobileNetV2 по всем разрешениям для различных градаций серого, усами обозначены 95% доверительные интервалы для матожидания;

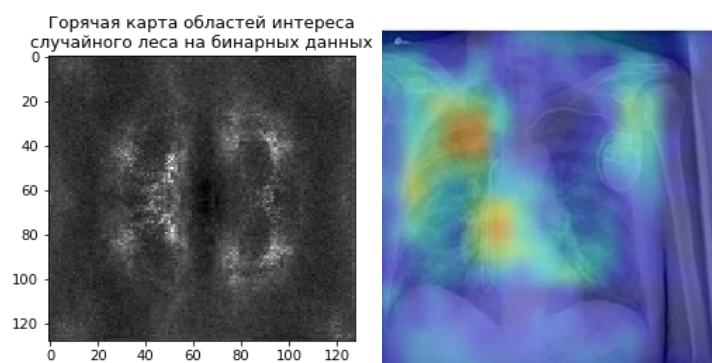


Рисунок 7. Горячая карта областей интереса случайного леса (слева), обученного на бинарных изображениях и карта GradCAM нейросети (справа)

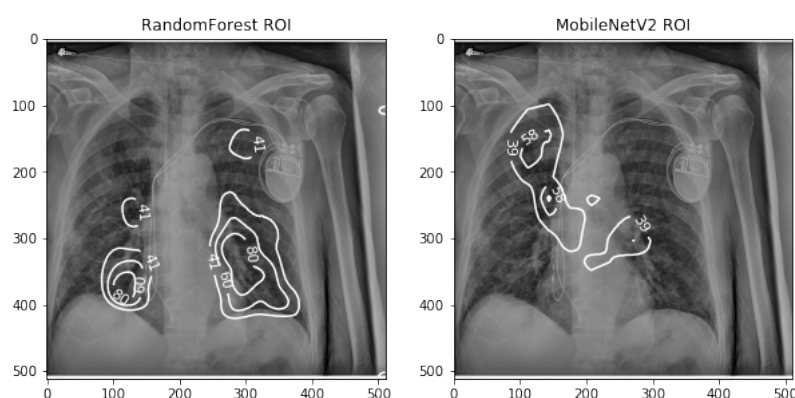


Рисунок 8. Области интересов обученных моделей классификаторов случайного леса (слева) и свёрточной нейронной сети (справа) для конкретного примера из набора данных

На рисунках 7, 8 представлены области интересов (ROI, Region of interest) для положительного диагностирования на пневмонию для двух видов моделей. Числами обозначаются проценты от максимального значения горячей карты. В качестве ансамбля деревьев решений использовалась модель, ранее обученная в полной цветности на разрешении 64 на 64 пикселей, а в качестве свёрточной нейронной сети MobileNetV2 выступила ранее обученная модель в полной цветности на разрешении 512 на 512 пикселей. Для получения ROI у случайного леса использовался путь решения каждого дерева с последующим усреднением по ансамблю. Для получения ROI у свёрточной нейронной сети использовался метод GradCAM [14], градиенты рассчитывались относительно выхода с последнего свёрточного

слоя размером 16 на 16 пикселей. Для получения более точных областей интереса, результаты усреднялись вместе с отраженными вдоль вертикальной оси изображениями. Для избавления от шумовой компоненты на полученный ROI случайного леса был наложен сглаживающий фильтр Гаусса с размером ядра, равным 11. Для наложения ROI на изображения с большим разрешением, разрешение областей интересов было увеличено при помощи бикубической интерполяции. Как видно из рисунка, для принятия решения о наличии пневмонии, области интересов случайного леса и дерева решений не совпадают. Случайный лес смотрит больше на нижние отделы лёгких, в то время как свёрточная нейронная сеть акцентирует внимание на верхнем отделе. Однако, существуют и общие области интересов, например средний отдел правого лёгкого с уровнем интереса в 41% у случайного леса и 58% у свёрточной нейронной сети. А также средний отдел левого лёгкого с областями интересов в 41% и 39% для ансамбля деревьев решений и MobileNetV2 соответственно. Оба классификатора верно классифицируют наличие пневмонии, согласно разметке из набора данных.

Были подогнаны три регрессионных модели методом наименьших квадратов на полученных экспериментальных данных (рисунок 2): линейная модель, полиномиальная модель 2ого порядка отклика поверхности (Response Surface Regression) и модель по сформированному ранее интегральному критерию (2). В таблице 1 представлены результаты для 3 моделей.

Таблица 1 – Результаты дисперсионного анализа для трёх регрессионных моделей

Параметр	Линейная	Отклика	Предложенная
R	0,938152	0,968969	0,965308
R^2	0,880129	0,938901	0,931819
Скорректированный R^2	0,879784	0,938723	0,931621
Степени свободы	2	5	4
Критерий Фишера	1273,894	1057,241	2371,196
p-значение	0,00	0,00	0,00

Как видно из результатов по скорректированному коэффициенту детерминации с учетом степеней свободы математической модели, предложенный интегральный критерий лучше, чем обычная линейная модель. Предложенный интегральный критерий оценивания сравним по качеству с полиномиальной моделью второго порядка при меньшем числе степеней свободы. По рассчитанному критерию Фишера гипотеза о нулевых значениях коэффициентов, т.е. об одновременном равенстве нулю всех коэффициентов кроме свободного члена, не принимается ни для какой из моделей с 95% статистической значимостью. Подобранные параметры для математических моделей представлены в таблицах 2-4. Полученные значения коэффициентов имеют статистическую значимость. Для всех трёх моделей влияние разрешения оказывает большее влияние на качество, нежели число градаций серого: разрешение в 9 раз имеет больший вес у линейной модели, в 5 раз у полиномиальной и в 9 раз у предложенной математической модели интегрального критерия качества. Полученная полиномиальная модель 2ого порядка отклика поверхности достигает максимума в 89,16% качества в точке (1,2316; 0,9553), что соответствует размеру изображения в 2171 пиксель и градаций серого, равным 2913. Если брать ближайшие кратные двойки, то получим 2048x2048 для разрешения и значения между 2048 и 4096

для цветности. Данные значения выходят за пределы исследованной экспериментальной области разрешения и цветности. Кроме того, наблюдаемое ухудшение качества при размере стороны изображения в 512 пикселей свидетельствует либо о локальном максимуме изученной поверхности, либо о том, что построенная полиномиальная модель некорректна.

Таблица 2 – Коэффициенты линейной модели

Параметр	Среднее	Стандарт. откл.	t-критерий Стьюдента	p-значение
Свободный член	0,569221	0,004159	136,8605	0,000000
Коэффициент при R_{norm}	0,298042	0,005944	50,1395	0,000000
Коэффициент при D_{norm}	0,034059	0,005857	5,8151	0,000000

Таблица 3 – Коэффициенты модели отклика

Параметр	Среднее значение	Стандартное отклонение	t-критерий Стьюдента	p-значение
Свободный член	0,536685	0,004824	111,2576	0,000000
Коэффициент при R_{norm}	0,503253	0,016512	30,4773	0,000000
Коэффициент при R_{norm}^2	-0,235637	0,015163	-15,5399	0,000000
Коэффициент при D_{norm}	0,094167	0,016007	5,8829	0,000000
Коэффициент при D_{norm}^2	-0,101360	0,014094	-7,1917	0,000000
Коэффициент при $R_{norm} \times D_{norm}$	0,080787	0,013159	6,1391	0,000000

Таблица 4 – Коэффициенты модели предложенного интегрального критерия

Параметр	Среднее значение	Стандартное отклонение	t-критерий Стьюдента	p-значение
β	0,519061	0,005326	97,46548	0,000000
α	0,450209	0,010498	42,88380	0,000000
a1	0,229160	0,013125	17,45994	0,000000
γ	0,052211	0,006040	8,64408	0,000000
a2	0,179034	0,070325	2,54582	0,011337

На рисунке 9 представлены трехмерные визуализации подобранных зависимостей для 3 регрессионных моделей.

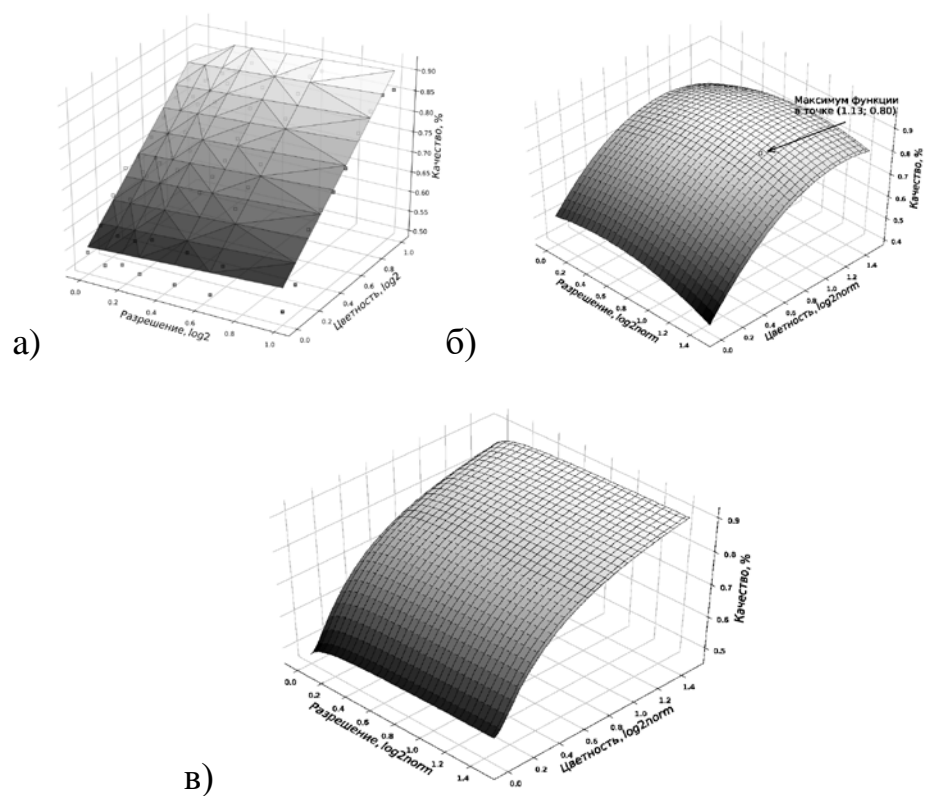


Рисунок 9. Поверхности регрессионных моделей: а) линейная модель; б) модель отклика поверхности; в) предложенная модель

Заключение

В курсовой работе было исследовано влияние цветности и разрешения рентенографических снимков и КТ на наборе данных SIIM-FISABIO-RSNA COVID-19. Проанализирован формат DICOM для представления медицинских данных, изучены его особенности представления пискельных данных.

Был сформулирован интегральный критерий качества снимков как математическая модель цветности, разрешения и систематической ошибки. Для оценки качества по предложенному интегральному критерию были обучены модели машинного обучения: ансамбли деревьев решений и свёрточной нейронной сети MobileNetV2 в различных конфигурациях изучаемых параметров. Для оценки качества моделей использовалась комплексная метрика бинарной классификации третьего уровня ROC-AUC. Для свёрточной нейросети было достигнуто качество $87,0 \pm 1,0\%$ при размере стороны изображения 256 пикселей и градаций серого, равным 256. Для случайного леса было достигнуто качество $77,0 \pm 1,5\%$ при полной цветности и разрешении 16 на 16 пикселей.

При помощи случайного леса для бинаризации снимков был получен оптимальный порог для разбиения, равный старшему биту цветности. При этом качество модели ухудшается лишь на 2%, что сравнимо с качеством MobileNetV2, обученному при высоких разрешениях и бинарных изображениях. По полученным экспериментальным данным, для получения адекватной свёрточной нейронной сети достаточно 16 градаций серого с размером стороны, равным 256.

Были проанализированы области интересов моделей ансамблей деревьев решений и свёрточной нейронной сети. Возможность получить данные области позволят помочь увеличить эффективность специалистов, занимающихся диагностикой больных SARS-Cov-2.

По полученным экспериментальным данным были подогнаны три математические регрессионные модели методом наименьших квадратов. По коэффициенту детерминации можно судить о высоком качестве предложенной математической модели изображений. Подобранные оценки коэффициентов регрессионных моделей позволяют судить о девятикратно большем влиянии размера снимка на качество, нежели его цветности.

Список использованной литературы

1. Электронные лекции по курсу «Машинное обучение» МГТУ им. Н.Э. Баумана (автор Ю.Е. Гапанюк)
2. Talha Anwar and Seemab Zakir. Deep learning based diagnosis of covid-19 using chest ct-scan images. In 2020 IEEE 23rd International Multitopic Conference (INMIC), pages 1– 5, 2020
3. Xiaohong Gao, Yu Qian, and Alice Gao. Covid-vit: Classification of covid-19 from ct chest images based on vision transformer models. arXiv preprint arXiv:2107.01682, 2021.
4. Dimitrios Kollias, Athanasios Tagaris, Andreas Stafylopatis, Stefanos Kollias, and Georgios Tagaris. Deep neural architectures for prediction in healthcare. *Complex & Intelligent Systems*, 4(2):119–131, 2018.
5. Dimitris Kollias, Y Vlaxos, M Seferis, Ilianna Kollia, Levon Sukissian, James Wingate, and Stefanos D Kollias. Transparent adaptation in deep medical image diagnosis. In TAILOR, pages 251–267, 2020.
6. Mildenberger P., Eichelberg M., Martin E. Introduction to the DICOM standard //European radiology. – 2002. – Т. 12. – №. 4. – С. 920-927.
7. Lakhani P, Mongan J, Singhal C, Zhou Q, Andriole KP, Auffermann WF, Prasanna P, Pham T, Peterson M, Bergquist PJ, Cook TS, Ferraciolli SF, de Antonio Corradi GC, Takahashi M, Workman SS, Parekh M, Kamel S, Galant JH, Mas-Sanchez A, Benítez EC, Sánchez-Valverde M, Jaques L, Panadero M, Vidal M, Culiáñez-Casas M, Angulo-Gonzalez DM, Langer

- SG, de la Iglesia Vaya M, Shih G. The 2021 SIIM-FISABIO-RSNA Machine Learning COVID-19 Challenge: Annotation and Standard Exam Classification of COVID-19 Chest Radiographs. [Internet]. OSF Preprints; 2021. Режим доступа [Электронный ресурс]: <https://doi.org/10.31219/osf.io/532ek>, Дата обращения: 12.12.2021
8. Tsai, E., Simpson, S., Lungren, M.P., Hershman, M., Roshkovan, L., Colak, E., Erickson, B.J., Shih, G., Stein, A., Kalpathy-Cramer, J., Shen, J., Hafez, M.A.F., John, S., Rajiah, P., Pogatchnik, B.P., Mongan, J.T., Altinmakas, E., Ranschaert, E., Kitamura, F.C., Topff, L., Moy, L., Kanne, J.P., & Wu, C. (2021). Data from Medical Imaging Data Resource Center (MIDRC) - RSNA International COVID Radiology Database (RICORD) Release 1c - Chest x-ray, Covid+ (MIDRC-RICORD-1c). The Cancer Imaging Archive. Режим доступа [Электронный ресурс]: <https://doi.org/10.7937/91ah-v663>, Дата обращения: 12.12.2021
 9. Prieto C. et al. Image retake analysis in digital radiography using DICOM header information //Journal of digital imaging. – 2009. – Т. 22. – №. 4. – С. 393-399
 10. Atkinson S., Neep M., Starkey D. Reject rate analysis in digital radiography: an Australian emergency imaging department case study //Journal of medical radiation sciences. – 2020. – Т. 67. – №. 1. – С. 72-79.
 11. Robertson S., Zaragoza H. The probabilistic relevance framework: BM25 and beyond. – Now Publishers Inc, 2009.
 12. Sandler M. et al. Mobilenetv2: Inverted residuals and linear bottlenecks //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2018. – С. 4510-4520.
 13. Kingma D. P., Ba J. Adam: A method for stochastic optimization //arXiv preprint arXiv:1412.6980. – 2014.
 14. Selvaraju R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization //Proceedings of the IEEE international conference on computer vision. – 2017. – С. 618-626.