

Оценка эпистемической неопределённости для свёрточных нейронных сетей в задаче определения агглютинации

Егоров Ф.Д.^{1,2}, Корчагин С.А.^{1,2}, Ершов Е.И.^{1,2}

¹ Институт проблем передачи информации им. А.А. Харкевича РАН, 127051,
г.Москва, Б. Картеный пер., д.19, стр.1

² Московский физико-технический институт (НИУ), 141701, Россия, Московская
область, г. Долгопрудный, Институтский пер., 9.
egorovf9@gmail.com

Аннотация В работе приводится распределение предельной точности 70 свёрточных нейронных сетей, независимо обученных для задачи бинарной классификации изображений круглых лунок со смесью крови и реагента по признаку наличия агглютинации. Гиперпараметры одинаковы для всех сетей, включённых в выборку. На основе полученного распределения определены его статистические параметры: среднее, медиана и дисперсия. Найдены статистические оценки оптимального количества сетей, которые необходимо обучить для получения точности, большей или равной заранее заданного значения. Дополнительно получены оценки для математического ожидания и стандартного отклонения в зависимости от размера подвыборки из обученных сетей.

Ключевые слова: Свёрточные нейронные сети, анализ изображений, математическая статистика, агглютинация, группа крови

1 Введение

Определение группы крови является одной из наиболее важных задач в медицине. Она связана с широким спектром процедур и исследований, начиная от определения предрасположенности к определённым заболеваниям и заканчивая переливанием крови и трансплантацией органов. Ошибка в определении группы крови перед переливанием может повлечь за собой смерть или серьёзный вред здоровью. Организм реципиента воспринимает кровяные клетки определённых групп как чужеродное тело, что может привести к тяжёлому состоянию, известному как гемолитический шок.

В настоящее время существует несколько способов определения групп крови, различающихся по стоимости, точности и количеству определяемых параметров. Наиболее простыми и распространёнными являются тестирование на агглютинацию (склеивание клеток крови) и гемолиз (разрушение клеток крови) при контакте с определёнными реагентами. Из более сложных

и подробных можно выделить молекулярно-генетическое исследование с использованием ПЦР (полимеразная цепная реакция) и масс-спектрометрию.

Исследование на агглютинацию, как правило, проводится лаборантом, и наличие реакции определяется эмпирически на основе зрительного анализа смеси образца и реагента. Рассматриваемый в работе метод предполагает использование легковесной свёрточной нейронной сети для обнаружения агглютинации на основе изображения лунки с исследуемым реакцией. Такой подход характеризуется невысокой стоимостью, позволяет лаборанту избежать ошибок при анализе, снижая риск человеческой ошибки, а также позволяет автоматизировать запись результатов для хранения в лабораторной системе [1].

Предельной точностью нейронной сети назовём такое значение точности (Accurasy), к которому стремится точность предсказаний на рассматриваемой выборке при увеличении числа эпох. Это позволяет рассуждать о предельных значениях в реальных случаях, где точность может зачастую отличаться от максимально возможной из-за нестабильного процесса обучения.

В данной работе предельная точность принимается равной точности на последней эпохи обучения. Это основано на предположении, что количество эпох выбрано правильно, и при продолжении обучения точность будут только колебаться около некоторой средней величины, которая и является предельной точностью. Влияние таких флуктуаций на измерения, в свою очередь, уменьшится за счёт большого количества измеренных сетей.

Несмотря на то, что обученные модели с выбранной архитектурой показывают высокую предельную точность, её значения могут отличаться на несколько процентов для отдельно взятых сетей. Кроме того, значения точности по эпохам имеют заметные случайные колебания на всех этапах обучения. Эти наблюдения говорят о наличии эпистемической неопределенности для рассматриваемого метода. В рамках исследования проводится обучение множества сетей с одинаковой архитектурой и параметрами, затем - построение распределения их предельной точности. На основе полученной статистики при помощи интервальных оценок было оценено число сетей, которые необходимо обучить для получения точности, имеющей заданную величину. Также при помощи теоремы теоремы Колмогорова приведены оценки для математического ожидания и стандартного отклонения в зависимости от числа нейронных сетей в подвыборке.

2 Набор данных

Все изображения из использованного набора данных состоят из 3139 квадратных, центрированных изображений лунок в высоком разрешении (каждая сторона - 512 пикселей), сделанных с помощью сканера (смотри Рисунок 1). Размер входного изображения выбран из соображений наилучшего соответствия размера, покрываемого одним пикселием, характерному размеру сгустка эритроцитов. Для каждой фотографии размечены, в частности,

признак агглютинации (присутствует или отсутствует), полученный из медицинских карт доноров крови.

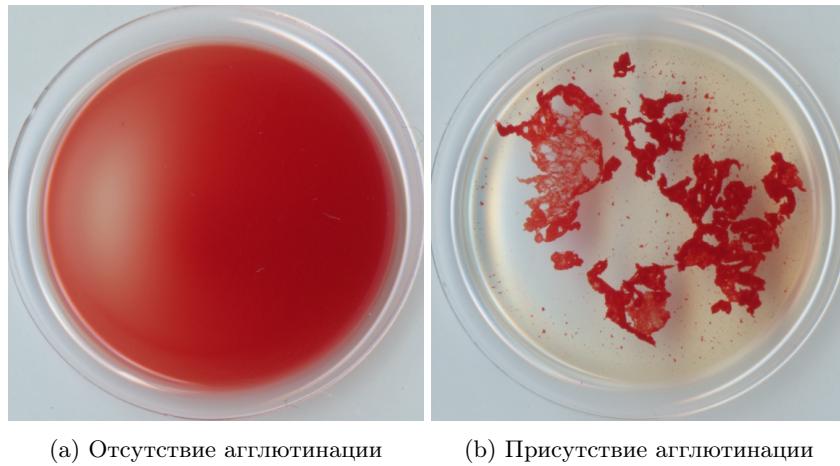


Рис. 1: Примеры изображений из набора данных.

3 Архитектура модели

Модель принимает на вход трёхканальное (RGB) изображение размера 512x512 и возвращает одно число $y \in [0; 1]$. Используемая архитектура описана на Рисунке 2 и состоит из семи последовательных слоёв: пяти свёрточных, с функцией активации ReLU (четыре первых из которых также дополняются MaxPool), и двух полносвязных с функциями ReLU и sigmoid соответственно. При обучении к первому полносвязному слою применяется Dropout с вероятностью 0.3.

Вне режима обучения в качестве итогового вердикта модели берётся ближайшее к y целое число:

$$f(y) = \begin{cases} 0 : y < 0,5 \\ 1 : y \geq 0,5 \end{cases} \quad (1)$$

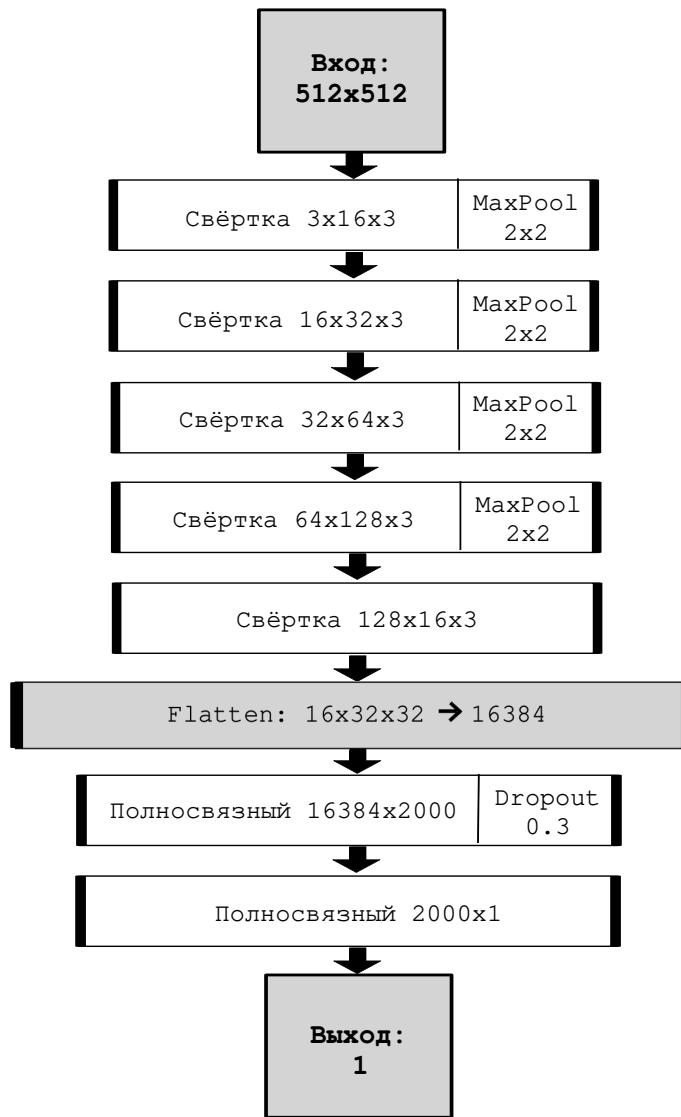


Рис. 2: Используемая архитектура.

4 Обучение

В ходе эксперимента независимо друг от друга были обучены 70 нейронных сетей, имеющих одинаковую архитектуру, гиперпараметры и тренировочный набор данных. Для обучения моделей были подобраны гиперпараметры, описанные в Таблице 1.

Таблица 1: **Финальные подобранные гиперпараметры.**

Параметр	Значение
Количество эпох	800
Начальное значение Learning Rate	$3 * 10^{-4}$
Изменение Learning Rate за эпоху, (в) раз	0,99
Коэффициент регуляризации (Weight decay)	$1 * 10^{-5}$

Подбор параметров был осуществлён экспериментально. Для каждого рассмотренного сочетания гиперпараметров был обучен отдельный набор из 4 сетей; Комбинации ранжировались по средней по соответствующему набору точности, измеренной на валидационной выборке. Если точности двух и более комбинаций оказывались на одном и том же полуинтервале вида $(0.5k\%; 0.5(k+1)\%]$, $k \in \mathbf{Z}$, из них рассматривались только те, дисперсия точности в наборе которых принимала минимальное значение.

Сети, обученные на этапе подбора гиперпараметров, не были включены в итоговую экспериментальную выборку. Точность на тестовой части набора данных не измерялась до завершения обучения всех сетей, составляющих выборку, и не использовалась для выбора гиперпараметров.

Перед подачей изображения на вход нейронной сети области на расстоянии от центра большем, чем 0.95 от радиуса лунки, заполняются чёрным цветом (#000000). При обучении изображения аугментируются следующим образом:

1. Поворот вокруг центра от 180° до -180° с вероятностью 100%;
2. Применение Гауссовского шума с $\sigma = 6$ к значению каждого канала с вероятностью 30%;
3. Параллельный перенос со смещением по обеим осям от -15 до 15 пикселей с вероятностью 30%. Образовавшиеся пустоты по краям заполняются чёрным #000000;

Для каждой сети обучение продолжалось одинаковое количество эпох; Learning Rate уменьшался в фиксированное число раз в конце каждой эпохи. В качестве оптимизатора был использован Adam. На каждой эпохе тренировочные изображения случайным образом перемешивались и аугментировались. Распределение набора данных по выборкам можно увидеть в Таблице 2.

Таблица 2: **Размеры используемых выборок**

Тип выборки	Количество изображений
Обучающая	1958
Валидационная	652
Тестовая	529

5 Распределения для различных выборок

На Рисунке 3 приведены гистограммы для всех трёх использованных выборок изображений, с размеченными на ними средним арифметическим (чёрным), медианой (зелёным) и стандартным отклонением (синим пунктиром, отложенным на соответствующем его величине расстоянии, по обе стороны от среднего), а также аппроксимации распределений нормальным (серым). В таблице 3 приведены статистики для каждой из полученных выборок.

Таблица 3: Параметры распределения на разных выборках.

Тип выборки	Средняя точность	Медиана	Стандартное отклонение
Тестовая	97.54%	97.57%	0.50%
Валидационная	98.09%	98.08%	0.44%
Обучающая	99.05%	99.03%	0.29%

Отметим, что на всех выборках распределение почти симметрично, так как среднее с малой погрешностью совпадает с медианой. Большая часть значений находится в окрестности среднего с полушириной, равной стандартному отклонению. Плотность распределения убывает почти монотонно при удалении от среднего значения.

Из этих наблюдений следует, что полученную статистику можно хорошо приблизить нормальным распределением. В этом можно убедиться интуитивно, сравнив совпадающие и несовпадающие площади гистограммы и приближающей её кривой нормального распределения. В последующих экспериментах искомые величины будут получены для двух распределений: выборочного (отмечено как "Тестовая выборка") и его нормального приближения.

6 Оценка с помощью доверительных интервалов

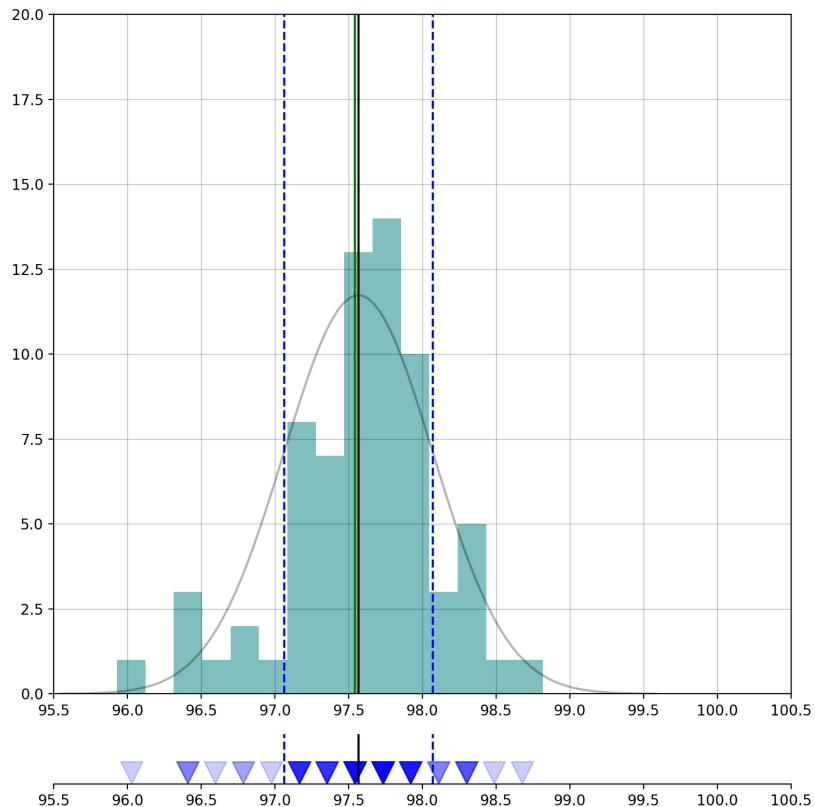
Одной из наиболее наглядных характеристик случайной величины являются доверительные интервалы. Задача построения доверительного интервала заключается в поиске множества \mathbf{I} некоторого вида (например, окрестности среднего), вероятность попадания в которое для рассматриваемой случайной величины x равна или превосходит некоторое установленное заранее значение C , называемое уровнем доверия. [3] Формально:

$$\mathbf{P}(x \in \mathbf{I}) \geq C. \quad (2)$$

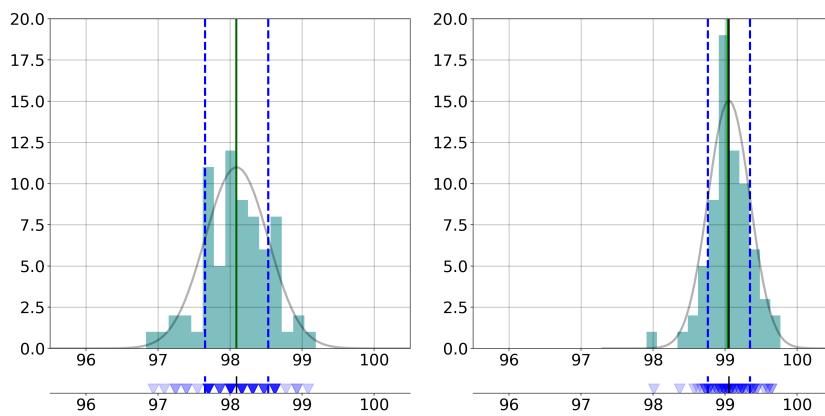
C наиболее часто принимают равным 90%, 95% или 99%.

Далее будем рассматривать предельную точность a нейронной сети как случайную величину, распределение которой известно приближённо по экспериментальным данным. Введём на её основе случайную величину

Рис. 3: Распределения точности на различных выборках.



(a) Тестовая выборка



(b) Валидационная выборка

(c) Обучающая выборка

$A_N = \max(a_1, \dots, a_n)$ — максимальную точность в наборе из N независимо обученных нейронных сетей. Теперь построим для A_N доверительный интервал в виде $[t, +\infty)$.

Для попадания A_N в доверительный интервал необходимо и достаточно, чтобы в него попала хотя бы одна из a_1, \dots, a_n . Примем вероятность того, что точность одной сети после обучения будет равна или превзойдёт t , отношению количества таких сетей в экспериментальной выборке $k_{acc \geq t}$ к её полному размеру k :

$$p = \frac{k_{acc \geq t}}{k}. \quad (3)$$

Теперь запишем вероятность случая, когда A_N не попадает в доверительный интервал, и выразим её же через C :

$$(1 - p)^N = \left(1 - \frac{k_{acc \geq t}}{k}\right)^N = 1 - C. \quad (4)$$

Обобщить формулу ещё больше не удаётся из-за невозможности преобразования $k_{acc \geq t}$ в числителе, поэтому точное определение $t(N)$ сильно зависит от распределения и неоднозначно.

Однако большую практическую ценность несёт решение обратной задачи — построения зависимости $N(t)$. По ней станет возможным оценить количество сетей, которые нужно обучить для получения желаемого уровня точности, тем самым найти оптимальное соотношение между максимизацией точности и минимизацией затрат на обучение.

Для этого возьмём логарифм от обеих частей равенства. Вероятность p теперь считается известной, так как t фиксирован и угадывать его по N подбором больше не требуется:

$$N = \log_{(1-p)}(1 - C) = \frac{\ln(1 - C)}{\ln(1 - p)} = \frac{\ln(1 - C)}{\ln\left(1 - \frac{k_{acc \geq t}}{k}\right)} \quad (5)$$

На Рисунке 4 изображён график полученной зависимости, а также оценка для нормального распределения. Она получается аналогично, но для получения p применены приближённые формулы для нормального распределения.

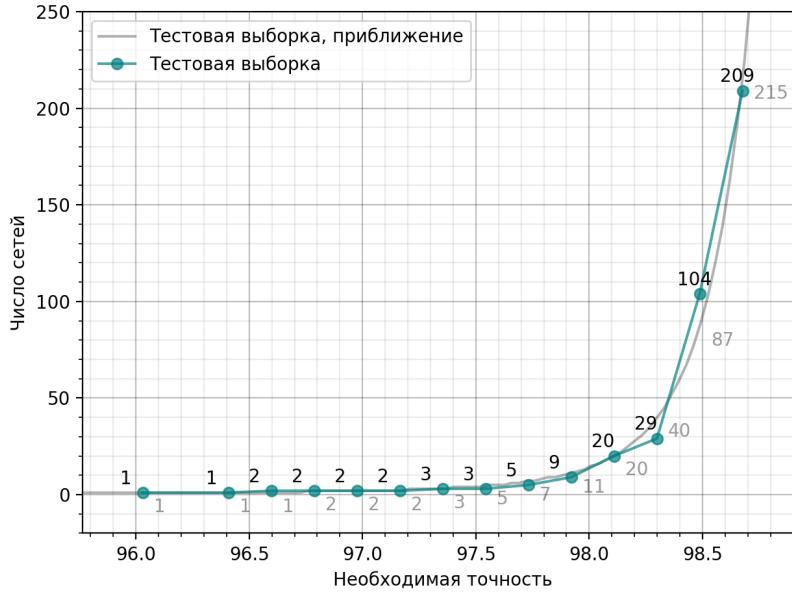


Рис. 4: Оценка оптимального количества сетей для уверенности 95%.

Для наборов из разного количества m сетей построим верхнюю оценку стандартного отклонения точности (Рисунок 5) и нижнюю оценку для её математического ожидания (Рисунок 6). Воспользуемся методом Монте-Карло: Для каждого M примем граничными соответствующие крайние значения матожидания и дисперсии, рассчитанные отдельно для каждого из 10000 случайных сочетаний без повторений, взятых из экспериментальной выборки:

$$inf(mean) = \min (mean(a_1^1 \dots a_m^1) \dots mean(a_1^{10000} \dots a_m^{10000})) \quad (6)$$

$$sup(std) = \max (std(a_1^1 \dots a_m^1) \dots std(a_1^{10000} \dots a_m^{10000})) \quad (7)$$

Где $mean$ используется для обозначения матожидания, std — стандартного отклонения. Надстрочные символы обозначают второй индекс вместо возведения в степень.

Иными словами, для каждого M находится худшая подвыборка и используется для оценки соответствующей величины.

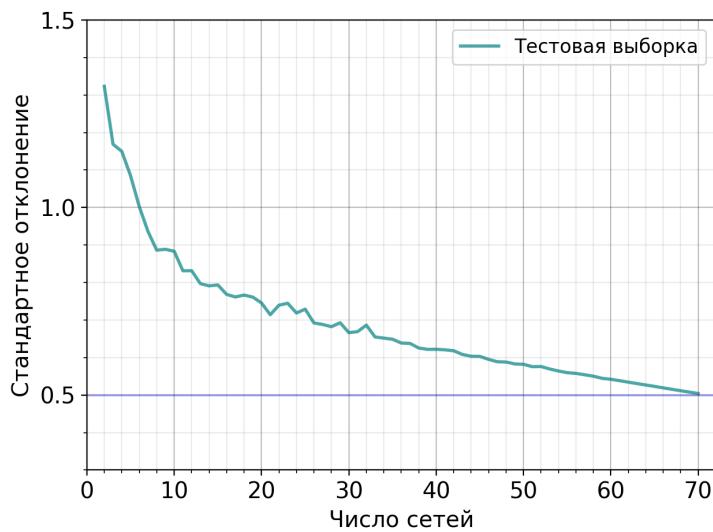


Рис. 5: Оценка максимального стандартного отклонения в наборе из M сетей. (Синим - значение на экспериментальной выборке)

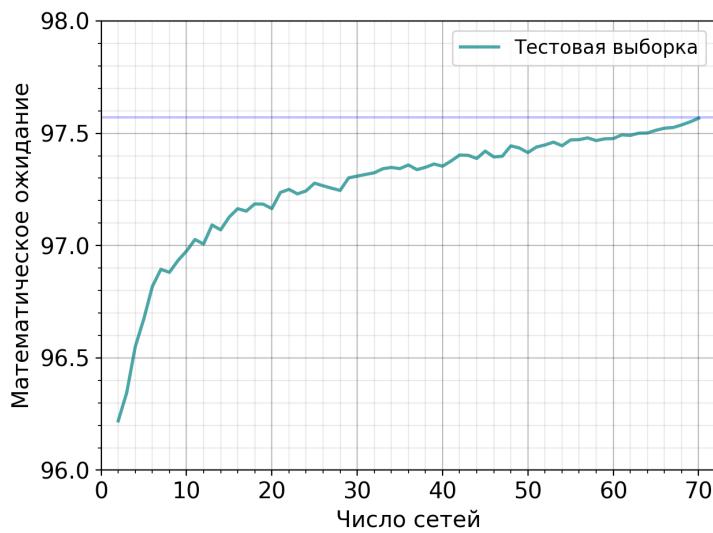


Рис. 6: Оценка минимального матожидания для набора из M сетей. (Синим - значение на экспериментальной выборке)

Отметим, что вид обеих кривых предполагает сходимость к теоретической величине. Это следует из равномерной сходимости функций распределений, построенных по конечным выборкам, к настоящему распределению, вытекающей из теоремы Колмогорова [2]. Это также обосновывает улучшение результатов в случае увеличения размера выборки.

7 Заключение

Исходя из изложенных выше результатов, в рамках выбранного уровня надёжности любая отдельно взятая сеть по точности превосходит среднестатистического эксперта, доля правильных предсказаний для которого 96.3% [1]. При пороге точности выше $\approx 98.3\%$ требуемое количество сетей резко возрастает, поэтому наилучшее соотношение между качеством и затратами на одну сеть достигается при обучении от 20 до 40 сетей. Для крайних оценок математического ожидания и стандартного отклонения аналогичный перегиб находится около 10 сетей, что можно рассматривать как минимальный размер набора.

Учитывая значительную величину стандартного отклонения по сравнению со средней долей неправильно данных моделью ответов (0.5% и 2.43%), можно судить о соизмеримости эпистемической неопределённости и альтернативной, оцененной по наивысшей полученной точности и соответствующей доле неправильных ответов (1.3%). Предположительно, использование более сложных архитектур может привести к лучшей точности и стабильности.

Список литературы

1. Корчагин, С.А. Алгоритм определения группы крови по изображениям серологических планшетов / С.А. Корчагин, Е.Е. Зайченкова, Д.А. Шарапов, Е.И. Ершов, Ю.В. Буторин, Ю.Ю. Венгеров // Компьютерная оптика. – 2023. (Проходит финальный процесс рецензирования)
2. Боровков А. А. Математическая статистика. — Санкт-Петербург : Лань, 2010. — С. 390. — ISBN 978-5-8114-1013-2.
3. Ш. Закс. Теория статистических выводов. — М.: Мир, 1975. — 776 с.