УДК 004.891.3

ОСОБЕННОСТИ ПРИМЕНЕНИЯ ПРЕДОБУЧЕННЫХ СВЁРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ К ЗАДАЧАМ СТЕГОАНАЛИЗА ГРАФИЧЕСКИХ ИЗОБРАЖЕНИЙ

© **С. Н. Терещенко**[1], **А. А. Перов**[2], **А. Л. Осипов**[1]

[1] *Новосибирский государственный университет экономики и управления, 630099, г. Новосибирск, ул. Каменская, 56*
[2] *Московский политехнический университет, 107023, Москва, ул. Б. Семёновская, 38*
E-mail: alosip@mail.ru
perov_artem@inbox.ru

Исследовано использование свёрточных нейронных сетей в целях анализа контейнера графических изображений на наличие данных, внедрённых методами стеганографии. Показано, что глубокая свёрточная нейронная сеть обучается классифицировать присутствие скрытых данных в графических изображениях, достигая точности по метрике weighted AUC, равной 0,928. Проверена гипотеза об эффективности применения концепции «transfer learning» в сфере стеганографии. Эффективность предложенной технологии продемонстрирована на большом экспериментальном материале.

Ключевые слова: машинное обучение, свёрточные нейронные сети, стеганография, стегоанализ, контейнер.

DOI: 10.15372/AUT20210412

Введение. Применение стеганографических методов защиты информации позволяет реализовать один из трёх базовых критериев информационной безопасности — конфиденциальность. Принципиальным отличием стеганографии от методов шифрования данных является то, что сама информация может не подвергаться прямому преобразованию и оставаться в своём первоначальном виде, однако скрывается сам факт передачи информации, которая внедряется в контейнер. Как правило, в современной интерпретации контейнером является элемент файловой системы, предназначенный для хранения скрываемой информации.

Идея внедрять секретную информацию в сообщения и контейнеры появилась ещё до нашей эры. Тогда одним из наиболее широко применяемых способов являлось использование специальных симпатических чернил, которые проявляли себя только под воздействием какого-либо химического реактива, освещения ультрафиолетом или термическим воздействием. Помимо проявляющихся чернил, применялись микроточки, содержащие сообщение в максимально физически сжатом виде, трафареты, являющиеся «картой» скрытого сообщения, и акростихи, в которых контейнер был устроен таким образом, что каждый символ слова становился составляющей секретного сообщения.

Одним из первых современных методов цифровой стеганографии [1, 2] является метод LSB (Least Significant Bit), основанный на внедрении сообщения в наименьшие значащие биты пикселей растрового изображения, что позволяет осуществить сокрытие в контейнере сообщения объёмом до $H \cdot W \cdot 3$ бит информации, где H — высота изображения, а W — ширина. Используя палитру RGB, пиксель контейнера может спрятать по биту в каждую из компонент цвета. Заполнение младших бит каждой из компонент цвета в пикселе не изменяет визуального восприятия цифрового изображения, однако даёт возможность

внедрить сообщение большого объёма. Такого рода методы позволяют использовать избыточность мультимедийных форматов (контейнером может служить не только цифровое, но аудио- и видеоизображение или текстовый файл). Цифровая стеганография использует глубокое обучение и нейронные сети. В [3] посредством применения свёрточных нейронных сетей производится внедрение информации в контейнер. Использование аппарата свёрточных нейронных сетей для внедрения скрытых сообщений в контейнеры, а также определения наличия такой скрытой информации в цифровых объектах (стегоанализ) представлено в [4–6], а в [7] предложен в качестве контейнера видеофайл с внедрённой в него информацией. Метод стеганографического преобразования двоичных сообщений, позволяющий встраивать скрытые данные с минимальным искажением его статистических свойств, продемонстрирован в [8].

Задачи по выявлению в контейнере внедрённых сообщений решаются другим научным направлением — стегоанализом. Методы анализа контейнеров с внедрённой информацией, основанные на сжатии данных, приведены в [9, 10]. В этих работах на примере графических изображений форматов BMP и JPEG, а также аудиофайлов WAV показано, что отрезки битовых последовательностей, не содержащих скрытой информации, сжимаются лучше, чем соответствующие им отрезки со стеговнедрёнными данными, что с высокой вероятностью может свидетельствовать о наличии внедрённой информации.

Применение математических методов стеганографии отражено в [11], где предлагается подход к стегоанализу с предварительной фильтрацией. Результаты экспериментов, проведённых в [11], показывают существенное снижение ошибки обнаружения внедрённой в контейнер информации. Алгоритмы обнаружения, основанные на математических методах, предложены в [12–14].

В настоящее время сфера применения технологий машинного обучения расширяется и открываются новые задачи. В частности, свёрточные нейронные сети всё чаще используются в задачах по информационной безопасности, что не является новым направлением. В [15] рассматривается применение технологий машинного обучения к задачам криптографии. Описывается известная в криптографии атака на побочные каналы (side-channel attack), а также упоминается несколько атак на криптоалгоритмы, основанные на технологиях машинного обучения. Предлагается атака на алгоритм AES, базирующаяся на глубоком обучении, которая является более эффективной, чем существующие шаблонные атаки. В [16] также рассматривается атака на побочные каналы, использующая технологии машинного обучения, цифровую подпись EdDSA, построенную на эллиптической кривой Эдвардса. В [17] предлагается универсальный метод статистического анализа, основанный на свёрточных нейронных сетях, который позволяет эффективно обнаруживать отклонения от случайности в выходных последовательностях итеративных алгоритмов шифрования на существенно меньших длинах выборок, чем в аналитических и статистических методах [18].

В [19] показано, что для классификации фрагментов гиперспектрального изображения очень эффективны предварительная трансформация его спектральных признаков к главным компонентам и последующее распознавание с помощью свёрточной нейронной сети, которая обучена на выборке, составленной из фрагментов этого изображения. Получены высокие проценты правильной классификации при работе с крупноформатным гиперспектральным изображением, причём часть классов, на которые разбито изображение, очень близка между собой и соответственно трудноразличима по гиперспектрам.

Одной из близких работ является [6], где рассматривается возможность использования аппарата свёрточных нейронных сетей для обнаружения стеганографических вложений в цифровых изображениях, разрабатывается своя модель на основе свёрточной нейронной сети. Результаты исследования демонстрируют возможность обнаружения до 85 % фактов

наличия стеганографических вложений с помощью достаточно простой в реализации модели, которая не использует сложных статистических алгоритмов. При построении модели нейронной сети можно применить два подхода. Первый предполагает разработку новой архитектуры нейронной сети и её обучение [6], второй (концепция transfer learning) — использование готовой нейронной сети, предобученной на глобальной базе изображений ImageNet, и переобучение последних слоёв на новой обучающей выборке.

Концепция transfer learning уже доказала свою эффективность во многих задачах. Целью предлагаемой работы является экспериментальное исследование возможности применения аппарата свёрточных нейронных сетей для обнаружения стеганографических вложений в цифровых изображениях. Рассмотрено применение архитектуры ResNet 50 к задачам обнаружения внедрённой информации в контейнер.

Материалы и методы исследования. В данной работе исходным материалом для исследований послужили размеченные изображения с внедрённой в контейнер информацией, а также без наличия таковой, опубликованные в открытом доступе платформы Kaggle [20]. Методы исследования: проектирование и разработка информационных систем, программирование, аугментация и расширение обучающей выборки для задач компьютерного зрения, настройка гиперпараметров обучения моделей нейронной сети для обработки графических изображений.

Результаты исследования. Открытая база обучающей выборки представляет собой 300 000 изображений, которые разделены на четыре класса: три класса с наличием вложенной информации методами трёх алгоритмов (JMiPOD, JUNIWARD, UERD) и один для обычного изображения. К набору изображений прилагается файл в формате CSV с разметкой фотографий. Для обучения использовался фреймворк PyTorch, библиотека torchvision. Решается задача классификации. Изображения размером 512×512 пикселей представлены в формате JPEG. Данный формат представляет собой алгоритм сжатия файла изображения, позволяющий уменьшить его размер без потери большого количества информации. Преобразование происходит поэтапно. Сначала изображение преобразуется в расширение YCbCr из каналов RGB. Затем DCT (дискретное косинусное преобразование) наносится на пиксели этих каналов с использованием соответствующих коэффициентов. Изображение, закодированное с помощью алгоритма JPEG, остаётся в цветовом пространстве YCbCr до тех пор, пока оно не будет декодировано программным обеспечением для просмотра изображений. Имеются подходы, которые дают возможность скрывать информацию в коэффициентах DCT различных каналов изображения JPEG, и поэтому полезная нагрузка (секретные данные) случайным образом распределяется между ними с учётом статистики коэффициентов DCT.

Изображения с низкой плотностью текстуры используются для сокрытия более коротких сообщений, а изображения с высокой плотностью текстуры — более длинных. Средняя длина сообщения составляет 0,4 бита на ненулевой коэффициент DCT.

Анализ различия по пикселям изображений со скрытой информацией, закодированных тремя алгоритмами, от эталонного изображения по R-каналу представлен на рис. 1, отклонение по DCT показано на рис. 2.

Таким образом, видно, что каждый из алгоритмов оставляет свои особые отличия в кодировании пикселей. Необходимо обучить нейронную сеть определять эти признаки. Для задачи классификации в работе была использована концепция transfer learning. Поскольку все современные нейронные сети используют анализ простейших графических примитивов на нижних слоях, то эта концепция зарекомендовала себя уже во многих задачах [21, 22]. Использование предобученной на большом количестве изображений нейронной сети с переобучением последних слоёв даёт преимущество в стоимости и скорости обучения. Предобученная модель начинает процесс обучения не с полного обучения всех слоёв, а с заданных паттернов обученных слоёв, которые были получены при решении другой задачи, сходной

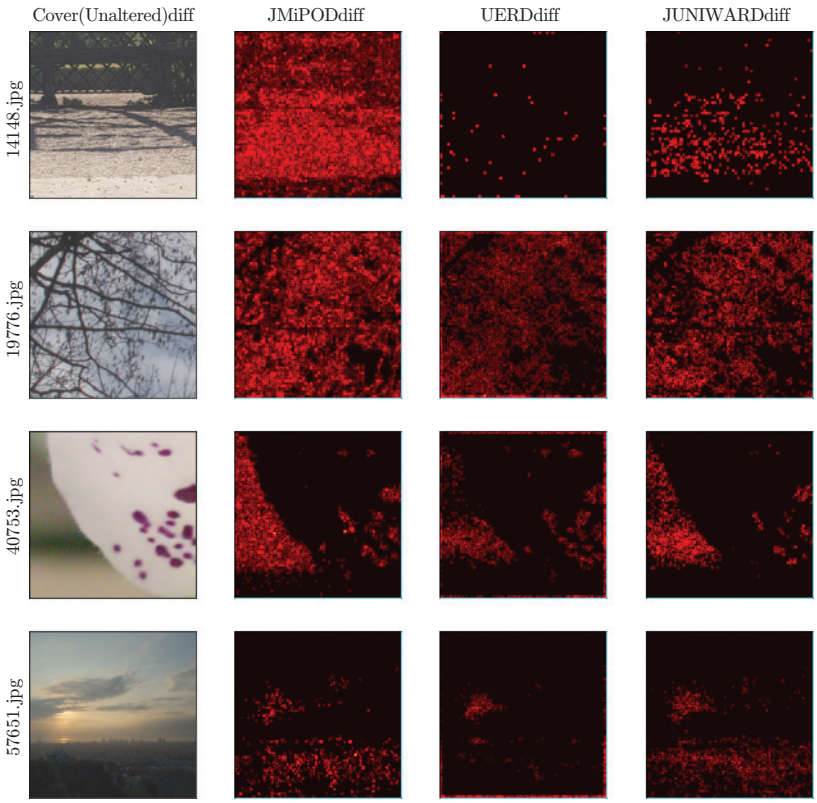

Рис. 1. Отклонение пикселей от эталонного изображения по R-каналу

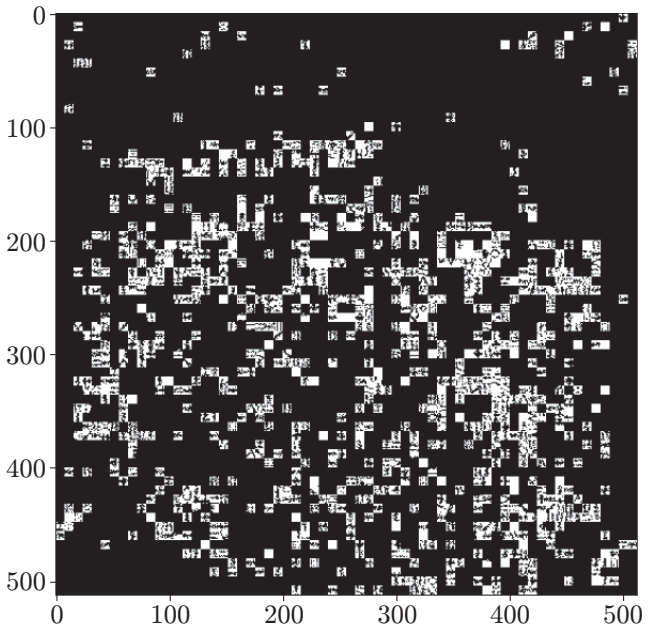

Рис. 2. Отклонение пикселей от эталонного изображения по DCT

Таблица 1

Архитектура ResNet 50

Стеки слоёв	Размер выхода	Слои
conv1	112×112	7×7, 64 / 2
conv2_x	56×56	3×3 max pool / 2
		$\begin{bmatrix} 1 \times 1, & 64, \\ 3 \times 3, & 64, \\ 1 \times 1, & 256 \end{bmatrix} \times 3$
conv3_x	28×28	$\begin{bmatrix} 1 \times 1, & 128, \\ 3 \times 3, & 128, \\ 1 \times 1, & 128 \end{bmatrix} \times 4$
conv4_x	14×14	$\begin{bmatrix} 1 \times 1, & 256, \\ 3 \times 3, & 256, \\ 1 \times 1, & 1024 \end{bmatrix} \times 6$
conv5_x	7×7	$\begin{bmatrix} 1 \times 1, & 512, \\ 3 \times 3, & 512, \\ 1 \times 1, & 2048 \end{bmatrix} \times 3$

по своей природе с решаемой. Предварительно обученная модель — это модель, которая была обучена на большом эталонном наборе данных (как правило, несколько десятков миллионов) для решения задачи, аналогичной классификации наличия дополнительной информации в изображении. Была выбрана предобученная модель ResNet 50.

В работе предложена гипотеза о том, что предобученная глубинная свёрточная нейронная сеть за счёт ключевой особенности анализа простейших примитивов изображения сможет выявить пиксельные отклонения для каждого из алгоритмов кодирования информации и изображения без таковых. Ключевой особенностью многослойной свёрточной нейронной сети ResNet является то, что она использует пропуск соединений или ярлыки для перехода через некоторые слои [23]. Типичные модели ResNet реализуются с двух- или трёхслойными пропусками, которые содержат нелинейности (ReLU) и пакетную нормализацию между ними. Сеть ResNet 50 является вариантом модели ResNet. В табл. 1 представлена архитектура ResNet 50. Нейронная сеть состоит из 50 слоёв, сгруппированных в 5 уровней.

В работе использовался метод стохастического градиентного спуска (SGD). В качестве настройки гиперпараметра шага обучения опытным путём подобрано значение $\mathrm{lr} = 0{,}003$. В качестве функции потерь была выбрана перекрёстная энтропия — мультиклассовая функция оценки логарифмических потерь, описанная в [24].

В качестве алгоритма стохастической оптимизации был использован алгоритм Adam. Общий набор изображений был разделён на три выборки: обучающую, проверочную и тестовую. Применялась предварительно обученная (на ImageNet) модель ResNet 50 с переобучением последних слоёв. В качестве метрики взята метрика weighted AUC (Area Under Curve) — площадь под кривой ошибок от (0,0) до (1,1) в координатах True Positive Rate (TPR) и False Positive Rate (FPR). Пороги заданы массивом [0,0; 0,4; 1,0]. Критерий AUC устойчив к несбалансированным классам и может быть интерпретирован как вероятность того, что случайно выбранный позитивный объект будет проранжирован классификатором выше (будет иметь более высокую вероятность быть позитивным), чем случайно выбранный негативный объект [25]. На рис. 3 представлены значения метрики weighted AUC.

Обучение состояло из 10 эпох. Как видно на рис. 4, на протяжении шести эпох про-

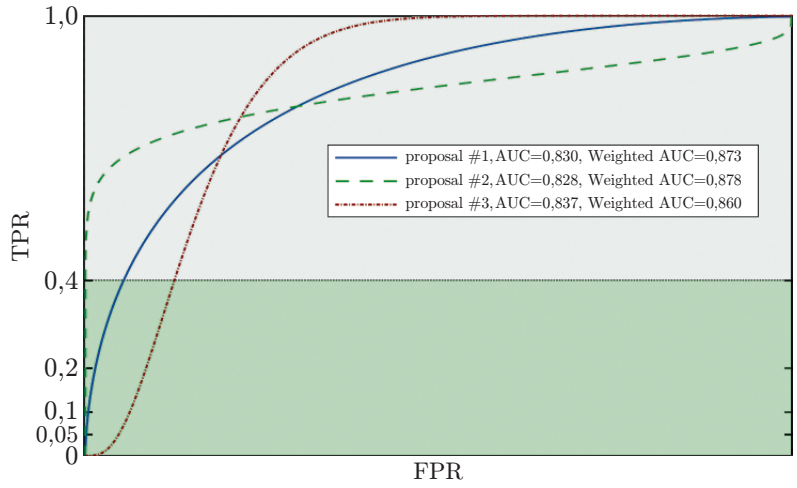

Рис. 3. Метрика weighted AUC [26]

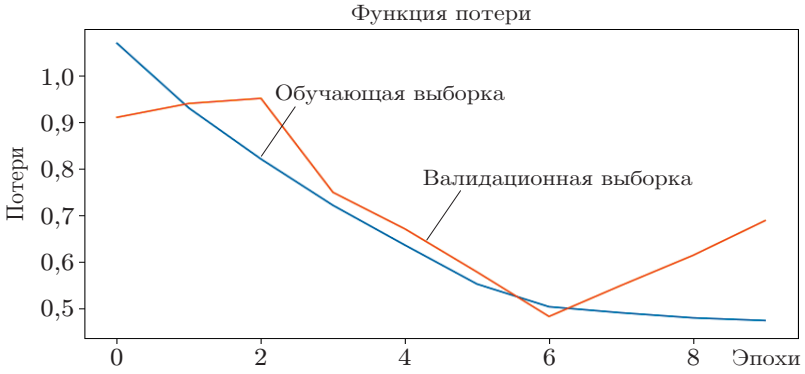

Рис. 4. Значения функции потерь при обучении модели

исходит снижение функции потерь синхронно по обучающей (трейн) и валидационной выборкам. После шестой эпохи наблюдается расхождение значений трейн- и валидационной выборок (первая продолжает уменьшаться, а вторая начинает повышаться), что свидетельствует о начале процесса переобучения нейронной сети.

Итоговые результаты, полученные по модели нейронной сети на тестовой выборке, показали, что точность классификации по метрике weighted AUC составила 0,928. Это позволяет сделать вывод о подтверждении гипотезы эффективности применения концепции transfer learning нейронных сетей для задач стеганографии.

Заключение. В данной работе исследован подход использования методов глубокого обучения для автоматической классификации и обнаружения скрытой информации в изображениях. Используя общедоступный набор данных из 300 000 изображений эталонных и закодированных по трём наиболее известным алгоритмам стеганографии, глубокая свёрточная нейронная сеть обучается классифицировать наличие скрытых данных, достигая точности по метрике weighted AUC, равной 0,928. В ходе исследований разработана модель свёрточной нейронной сети для обнаружения факта применения стеганографии в цифровых изображениях, а также на основании созданной модели реализована программа для стегоанализа, реализующая классификацию цифровых изображений. Разработанная модель показала более высокую точность по сравнению со сложными статистическими классификаторами. К достоинствам предложенного способа обнаружения стеговложений можно отнести высокую точность и простоту реализации. Направлением дальнейших исследований является совершенствование созданной модели с целью повышения вероятно-

сти обнаружения стеганографических вложений и снижения вычислительной сложности её программной реализации, а также создание собственной базы изображений с различными параметрами вложений.

СПИСОК ЛИТЕРАТУРЫ

1. **Bhaskari L., Avadhani P., Damodaram A.** Watermark insertion Algorithm implementation using Auxiliary carry and LSB methods // Proc. of the Intern. Conference on Systemics, Cybernetics and Informatics. Hyderabad, India, 8 Jan., 2006. P. 666–668.
2. **Zhang W., Zhang X., Wang S.** A double layered "plus-minus one" data embedding scheme // IEEE Signal Proc. Lett. 2007. **14**, N 11. P. 848–851.
3. **Kumar V., Laddha S., Aniker D. N.** Steganography techniques using convolutional neural networks // Rev. Comput. Eng. Studies. 2020. **7**. P. 66–73.
4. **Husien S., Badi H.** Artificial neural network for steganography // Neural Comput. & Applic. 2015. **26**. P. 111–116. DOI: 10.1007/s00521-014-1702-1.
5. **Sharma A., Kushwaha A.** Image steganography scheme using neural network in wavelet transform domain // Intern. Journ. Science and Research (IJSR). 2015. **4**, Iss. 12. P. 1773–1778.
6. **Полунин А. А., Яндашевская Э. А.** Использование аппарата свёрточных нейронных сетей для стегоанализа цифровых изображений // Тр. ИСП РАН. 2020. **32**, Вып. 4. С. 155–164. DOI: 10.15514/ISPRAS-2020-32(4)-1.
7. **Velmurugan K., Subramani H.** Video steganography by neural networks using Hash function // Proc. of the 5th Intern. Conference on Science Technology Engineering and Mathematics (ICONSTEM). Chennai, India, 14–15 March, 2019. P. 55–58. DOI: 10.1109 / ICONSTEM.2019.8918877.
8. **Нечта И. В.** Метод стеганографического преобразования сообщения со свойством частичной неизвлекаемости // Вычислительные технологии. 2019. **24**, № 3. С. 75–87.
9. **Жилкин М. Ю.** Стегоанализ графических данных на основе методов сжатия // Вестн. СибГУТИ. 2008. № 2. С. 62–66.
10. **Забелин М. А.** Стегоанализ аудиоданных на основе методов сжатия // Вестн. СибГУТИ. 2010. № 1. С. 41–49.
11. **Монарёв В. А., Пестунов А. И.** Повышение эффективности методов стегоанализа при помощи предварительной фильтрации контейнеров // Прикладная дискретная математика. 2016. № 2(32). С. 87–99.
12. **Yang Z., Huang Y., Zhang Y.** A fast and efficient text steganalysis method // IEEE Signal Process. Lett. 2019. **26**, Iss. 4. P. 627–631. DOI: 10.1109/LSP.2019.2902095.
13. **Jin Z., Feng G., Ren Y., Zhang X.** Feature extraction optimization of JPEG steganalysis based on residual images // Signal Processing. 2020. **170**. 107455. DOI: 10.1016/j.sigpro.2020.107455.
14. **Soto R. T., Ramos-Pollan R., Isaza G. et al.** Digital Media Steganalysis // Digital Media Steganography: Principles, Algorithms, and Advances /Ed. M. Hassaballah. Academic Press, 2020. 386 p. DOI: 10.1016/B978-0-12-819438-6.00020-7.
15. **Alani M. M.** Applications of machine learning in cryptography: A survey // Proc. of the 3rd Intern. Conference on Cryptography, Security and Privacy. Kuala Lumpur, Malaysia, 19–21 Jan., 2019. P. 23–27. DOI: 10.1145/3309074.3309092.
16. **Weissbart L., Picek S., Batina L.** One trace is all it takes: Machine Learning-based Side-channel Attack on EdDSA / Cryptology ePrint Archive: Report 2019/358, 2019. 18 p.
17. **Перов А. А., Пестунов А. И.** О возможности применения свёрточных нейронных сетей к построению универсальных атак на итеративные блочные шифры // Прикладная дискретная математика. 2020. № 3 (49). С. 46–57.

18. **Osipov A. L., Bobrov L. K.** The use of statistical models of recognition in the virtual screening of chemical compounds // Automatic Documentation and Mathematical Linguistics. 2012. **46**, N 4. P. 153–158.
19. **Козик В. И., Нежевенко Е. С.** Классификация гиперспектральных изображений с помощью свёрточных нейронных сетей // Автометрия. 2021. **57**, № 2. С. 13–21. DOI: 10.15372/AUT20210202.
20. **Kaggle.** Система организации конкурсов по исследованию данных, а также социальная сеть специалистов по обработке данных и машинному обучению. Kaggle Inc., 2021. URL: www.kaggle.com. (дата обращения: 05.04.2021).
21. **Rahman C. R., Arko P. S., Ali M. E. et al.** Identification and recognition of rice diseases and pests using convolutional neural networks // Biosystems Engineering. 2020. **194**. P. 112–120. DOI: 10.1016/j.biosystemseng.2020.03.020.
22. **Karmokar B. C., Ullah M. S., Siddiquee Md. K., Alam K. Md. R.** Tea leaf diseases recognition using neural network ensemble // Intern. Journ. Comput. Appl. 2015. **114**, N 17. P. 27–30.
23. **He K., Zhang X., Ren Sh., Sun J.** Deep residual learning for image recognition // Comp. Vis. Pattern Recogn. Cornell Univers., 2015. URL: https://arxiv.org/abs/1512.03385 (дата обращения: 05.04.2021).
24. **Kim Y., Lee Y., Jeon M.** Imbalanced image classification with complement cross entropy // Comp. Vis. Pattern Recogn. Cornell Univers., 2021. URL: https://arxiv.org/pdf/2009.02189.pdf (дата обращения: 05.04.2021).
25. **Метрики** в задачах машинного обучения. «Habr», 2006–2021. URL: https://habrcom/ru/company/ods/blog/328372 (дата обращения: 05.04.2021).
26. **Alaska2 Image Steganalysis.** Detect secret data hidden within digital images. URL: https://www.kaggle.com/c/alaska2-image-steganalysis/overview/evaluation (дата обращения: 05.04.2021).

Поступила в редакцию 05.04.2021
После доработки 05.05.2021
Принята к публикации 10.05.2021