Инженерный вестник Дона, №2 (2025)
ivdon.ru/ru/magazine/archive/n2y2025/9859

Применение сверточных нейронных сетей и алгоритмов глубокого обучения для прогнозирования и идентификации голосовых дипфейков

К.Г. Пономарёв, Е.А. Верещагина

Институт математики и компьютерных технологий Дальневосточного федерального университета, Владивосток

Аннотация: Целью данной статьи является создание модели сверточной нейронной сети идентификации и прогнозирования аудио-дипфейков путем классификации голосового контента с применением алгоритмов глубокого машинного обучения, библиотек языка программирования «python». Наборы данных аудиоконтента являются базовыми для процесса обучения нейронной сети и представлены мел-спектрограммами. Обработка графических изображений аудиосигнала в формате тепловой карты формируют базу знаний сверточной нейронной сети. Результаты визуализации мел-спектрограмм в соотношении величины измерения частоты звука и мела определяют ключевые характеристики аудиосигнала и обеспечивают процедуру сравнения между реальным голосом и искусственной речью. Современные синтезаторы речи используют комплексную подборку и ведут формирование синтетической речи на основании записи голоса человека и языковой модели. Отметим значимость мел-спектрограмм, в том числе, для моделей синтезации речи, где данный вид спектрограмм используется для записи тембра голоса и кодировки оригинальной речи говорящего. Сверточные нейронные сети позволяют автоматизировать обработку мел-спектрограмм и выполнить классификацию голосового контента: оригинальный или фейковый. Проведенные эксперименты на тестовых голосовых наборах доказали успешность обучения и применения сверточных нейронных сетей, использующих изображения мел-кепстральных коэффициентов MFCC, для классификации и исследования аудио контента, и применения данного вида нейронных сетей в области информационной безопасности для выявления аудио дипфейков.

Ключевые слова: нейронные сети, выявление голосовых дипфейков, информационная безопасность, модели синтезирования речи, глубокое машинное обучение, категориальная кросс-энтропия, функция потерь, алгоритмы выявления голосовых дипфейков, сверточные нейронные сети, мел-спектрограммы

Введение. В последние годы большую популярность завоевали цифровые сервисы синтезатора речи, имитирующие голос популярного актера, политика или диктора в развлекательных целях. Генерация речи формируется нейронной сетью на основании оригинальной звуковой дорожки говорящего человека путем применения нейронной модели речи, а также с применением метода «text-to-speech» языковой модели произношения. Современные модели генерации позволяют разделить голосовую дорожку на семантические и акустические данные, определить

тембр говорящего человека, а также языковую модель и обеспечить генерацию семантически правдоподобных подходящих словосочетаний.

Рассмотрим многоязычную модель синтезатора речи «CosyVoice2», формирующие языковые токены, которые являются базовыми для процедуры генерации речи и вносят семантические словосочетания языкового произношения [1]. В данную модель встроена языковая текстовая мультимодальная модель (схема), позволяющая дополнить записанный голос человека фонемами или условными единицами языка произношения. На рисунке 1 представлена типовая логическая схема синтезирования голосового произношения за счет использования всех методов: голосовой записи, тембра голоса и фонем.

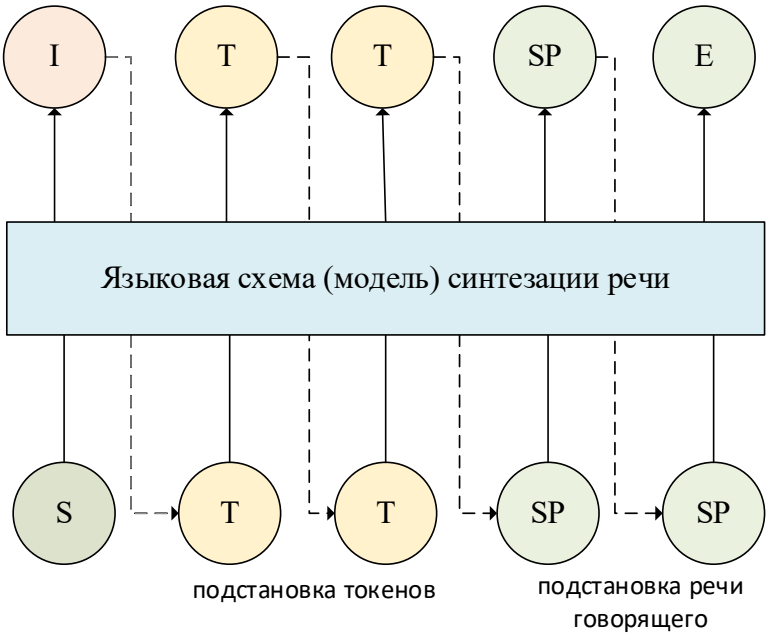

Рис. 1. – Логическая модель работы синтезатора речи с применением языковой схемы на примере модели «CosyVoice2»

где S – начало генерации речи, T – текстовый токен (фонема), I – исключение включения токена (фонема) или звуковой записи, SP – подстановка звуковой записи из выборки, E – конец генерации записи. Отметим, что действие исключения происходит при отсутствии логической целесообразности включения отрывка записи или текста, а также одновременно происходит

подбор подходящего словосочетания. При генерации синтетической речи используется метод конечного скалярного квантования, позволяющий улучшить правдоподобность произношения за счет исключения традиционного векторного квантования исходного сигнала. Граница эффективности скалярного квантования представлена функцией скорости искажения источника [2] по формуле нижней оценки границы Шеннона (1):

$$H(D) \geq H_{SH}(D) = H_0(X) - \frac{1}{2}\log_2(2\pi eD) \quad (1)$$

где D – среднеквадратичная ошибка, $H_{SH}(D)$ - граница Шеннона, $H_0(X)$ - относительная энтропия источника, вычисляемая по формуле (2):

$$H_0(X) = -\int_{-\infty}^{+\infty} f(X)\log_2 f(x)dx \quad (2)$$

где f(x) – функция плотности вероятности источника.

Генерация синтетической речи модели «CosyVoice2» по методу конечного скалярного квантования позволяет достичь оценочного балла распознавания речи «Mean Opinion Score» до 4,5 балла из 5 возможных. На информационном ресурсе «CosyVoice 2» по ссылке funaudiollm.github.io/cosyvoice2/ разработчики данного синтезатора речи предлагают ознакомиться и сравнить синтетическую и оригинальную речь говорящего.

В результате изучения современных синтезаторов речи сделан вывод, что синтетический голос максимально правдоподобен и использует логически правильные единицы лингвистики и языка произношения. Сгенерированный аудио дипфейк практически не отличается от оригинального произношения владельца голоса и требует специализированных программных средств выявления. Провели эксперимент по обучению нейронной сети на основании полученных критериев работы

ранее изученного синтезатора речи и взяли за основу обучения графические изображения мел-частотных кэпстральных коэффициентов MFCC для математического сравнения реальных и фейковых записей. Применили методы глубокого обучения сверточной нейронной сети, основанных на функции потерь и категориальной кросс-энтропии.

Постановка задачи исследования. Необходимо определить вид нейронной сети и метод математического выявления аудио дипфейков. Задача состоит в определении категории аудио контента путем исследования аудио сигнала через графическое изображение. Архитектура сверточной нейронной сети успешно проводит обработку изображения по формату яркости и оттенков, которые направлены в качестве входных данных мел-кепстральных спектрограмм MFCC и используют методы глубокого обучения [3]. Точность модели определяет порог ошибочных решений, при которых модель нейронной сети требует дополнительного обучения. Скрытые слои нейронной сети на каждом из слоев обучаются точно распознавать изображения. Операция свертки двумерного изображения I и ядра K представлена по формуле (3):

$$s(i,j) = \langle I * K \rangle (i,j) = \sum_m \sum_n I(m,n) K(i-m, j-n) \quad (3)$$

где I – двумерное изображение m*n, K - ядро. Данный подход позволяет более точно распознавать изображения и построить нейроны в каждом из слоев [4].

Решение задачи в условиях необходимости классификации входных данных. На первом этапе формируются графические изображения аудиосигнала мел-спектрограмм для обучения сверточной нейронной сети [5]. Преобразование исходного аудиосигнала проводится через дискретное преобразование Фурье по формуле (4):

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} = \sum_{n=0}^{N-1} x_n (\cos(2\pi kn / N) - i * \sin(2\pi kn / N) \quad (4)$$

где N - количество значений сигнала, измеренных за период, а также количество компонент разложения, X_n, n = 0,….., N-1 – измеренные значения сигнала в дискретных временных точках, X_k, k = 0,…., N-1, -N комплексных амплитуд синусоидальных сигналов. Для обучения сверточной нейронной сети воспользуемся библиотекой знаний «Kaggle» и выберем набор данных «The Fake-or-Real», имеющих в базе знаний 195 000 записей синтетической и реальной речи. Загрузку обучающей выборки в модель нейронной сети осуществляем через функцию keras библиотеки машинного обучения «TensorFlow 2.0» предварительно подготовив zip-архив аудиозаписей tf.keras.utils.get_file('real-or-fake.zip', origin=_URL, extract=True). Подготовка обучающих выборок проводилась по 2 категориям: реальная и фейковая запись. На рисунке 2 приведен пример генерации изображений с мел-частотными кэпстральными коэффициентами MFCC для определения характеристик реальной и синтетической речи.

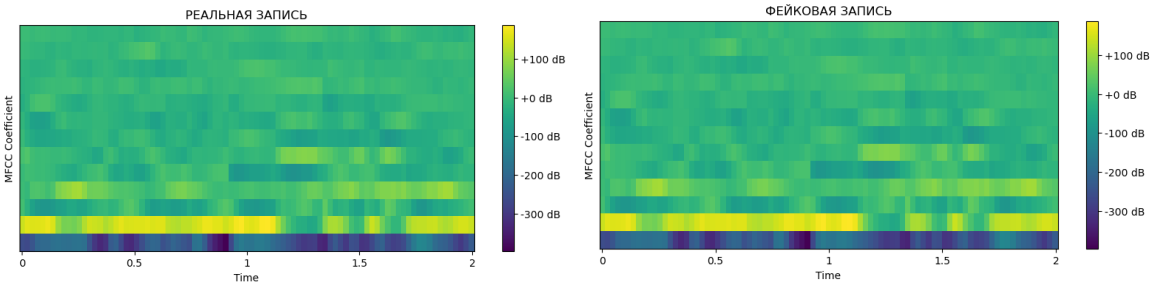

Рис. 2. – Пример формирования представления мел-частотных кэпстральных коэффициентов реальной и фейковой записи

Подготовлено 2 тестовых датасета. Аудио данные, полученные из библиотеки знаний, преобразованы в мел-частотные изображения. С помощью функции ImageDataGenerator(rescale=1./255) выполнены масштабирование изображений с коэффициентом 1/255.

В результате проведенного анализа доказана возможность визуального отображения аудиосигнала и формирования большой базовой выборки для обучения нейронной искусственной сети. Главным критерием совместимости мел-спектрограмм для взаимодействия нейронных сетей является точный формат входных изображений. Например, размер 28x28 пикселей по каждому изображению мел-спектрограммы, которые являются исследовательским инструментом аудио сигналов и могут сформировать базу знаний нейронной сети [6].

На рисунке 3 представим результаты построения модели нейронной сети и приведем функциональную блок-схему модели сверточной нейронной сети для поставленной задачи идентификации аудио дипфейков. Показаны следующие блоки: 1) входные параметры (мел-спектрограммы установленного размера, первичный входной слой нейронной сети); 2). Сверточная нейронная сеть (два сверточных слоя с нелинейной функцией активации «Retified Linear Unit» (RELU), слой выравнивания, скрытый слой с повторной активацией); 3). Выходной слой с двумя нейронами определяет истинность или фальсификацию аудиосигнала, разделив их на два класса: реальный и фейковый.

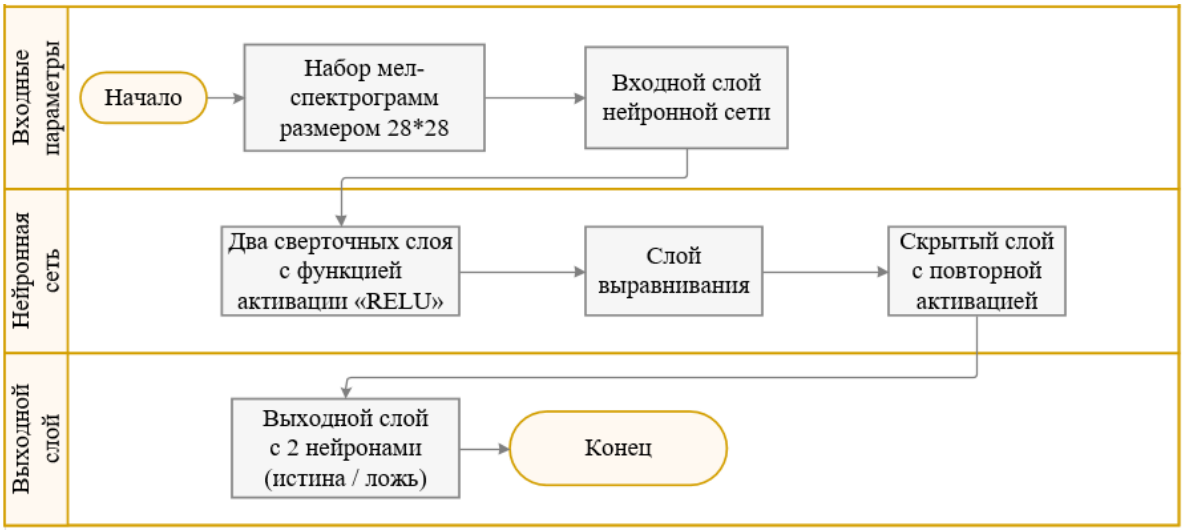

Рис. 3. – Функциональная блок-схема сверточной нейронной сети для выявления аудио-дипфейков

Для процесса обучения нейронной сети отметим значимость функции выпрямления и активации математической функции «RELU», где значения могут быть положительными и стремиться к бесконечности. Данная функция пропускает нейроны с положительными значениями и исключает нейроны с отрицательными значениями. Значимой функцией для обучения нейронной сети является алгоритм оптимизации. В языке программирования «python» используем функцию «Adam», которая позволяет сравнить весовые коэффициенты на основании входных и выходных данных. Процесс обучения строится на минимизации потерь или на сокращении ошибок в выходных данных. Функция потерь обновляет данные для увеличения точности. Кроме того, используем функцию категориальной кросс-энтропии, позволяющая на множестве категорий принять решение нейронной сетью к какой категории принадлежит входной аудио сигнал [7]. Формула категориальной кросс-энтропии представлена следующим видом (3)

$$H(y,\hat{y}) = -\sum_i y_i \log(\hat{y}_i) H(y,\hat{y}) = -\sum_i y_i \log(\hat{y}_i) H(y,\hat{y}) = -\sum_i y_i \log(\hat{y}_i) \quad (3)$$

Создана модель для обучения сверточной нейронной сети model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy']). Для обучения модели использована функция model.fit фреймворка «TensorFlow 2.0» [8]. На последнем полносвязном слое используем исключение результатов со значением вероятности 0.5 через функцию Dropout, которая представлена tf.keras.layers.Dropout (0.5). Вероятностные значения менее 50 %, поступающие на вход данного слоя, будут сброшены до 0. Таким образом, достигнута защита от переобучения. Создана последовательная модель и выполнена тренировка сверточной нейронной сети с помощью класса Sequential и сформирован стек последовательных слоев:

model = tf.keras.models.Sequential([tf.keras.layers.Conv2D(32, (3,3), activation='relu', input_shape=(IMG_SHAPE, IMG_SHAPE, 3)),

```
tf.keras.layers.MaxPooling2D(2, 2),
tf.keras.layers.Conv2D(64, (3, 3), activation='relu'),
tf.keras.layers.MaxPooling2D(2, 2),
tf.keras.layers.Conv2D(128, (3, 3), activation='relu'),
tf.keras.layers.MaxPooling2D(2, 2),
tf.keras.layers.Conv2D(128, (3, 3), activation='relu'),
tf.keras.layers.MaxPooling2D(2, 2),
tf.keras.layers.Flatten(),
tf.keras.layers.Dense(512, activation='relu'),
tf.keras.layers.Dense(2, activation='softmax')])
```

Функцией softmax разделены результаты значений на 2 класса: реальная или фейковая запись. Для обучения модели установим 90 эпох, в том числе определим количество тренировочных изображений для обработки перед обновлением параметров модели переменной batch_zise=100. Переменной total_train=1000 определим общее количество тренировочного набора данных, total_validate=2000 определим валидационный набор данных. Получена модель model.fit(train_data_gen, steps_per_epoch=int(np.ceil(total_train/float(batch_size))), epochs=90, validation_data=val_data_gen, validation_steps=int(np.ceil(total_validate/float(batch_size)))). Функция fit обеспечит обучение модели по установленным параметрам за одну эпоху. На рисунке 4 показаны полученные результаты обучения сверточной нейронной сети.

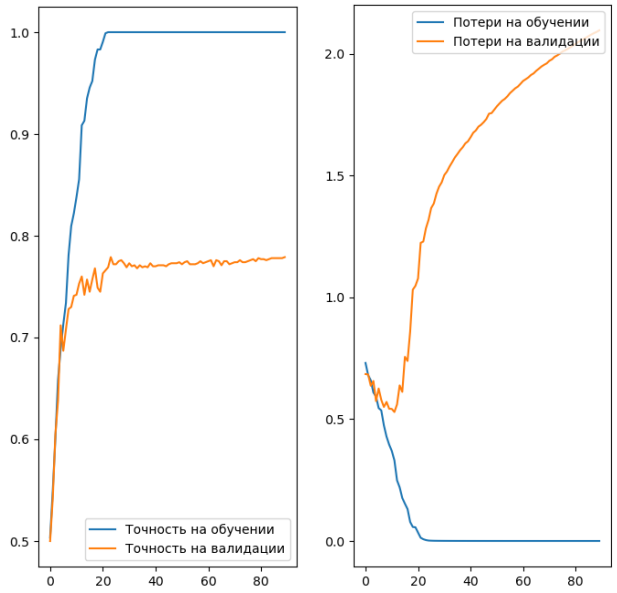

Рис. 4. – Точность и потери обучающих и валидационных данных

Проведенный эксперимент установил, что созданная модель сверточной нейронной сети с применением фреймворка «Tensorflow 2.0» [9] достигает точность вычислений в 78 %. Особенность созданной нейронной сети является использование графических изображений мел-кепстральных коэффициентов MFCC для выявления аудио дипфейков. На таблице 1 выполним сравнение между популярными моделями сверточных нейронных сетей для классификации графических изображений и созданной модели [10].

Таблица 1

Теоретическая точность выявления аудио дипфейков сверточных нейронных сетей

Наименование сети	Точность, %	Время обучения, сек.
ResNet-18 (90 эпох)	81,3 %	15127
FCNN-3 (90 эпох)	77,1 %	1011
Созданная модель (90 эпох)	78 %	977

Выводы. Доказана возможность применения сверточных нейронных сетей с использованием мел-кэпстральных коэффициентов MFCC для корреляции между реальной и фейковой записью, что позволило с точностью пока что 78 % идентифицировать синтетическую речь на небольшой обучающей выборке. Ранее нами были исследованы теоретические основы работы с мел-спектрограммами, основные характеристики и способы математического представления аудиосигнала [11]. В данной статье создана рабочая модель сверточной нейронной сети, результаты работы которой показали необходимость дальнейшего исследования по повышению точности выявления аудио дипфейков. В дальнейшем нами предполагается исследовать гармонические искажения и артефакты аудиосигнала, что позволит повысить показатель точности выявления аудио-дипфейков в предлагаемой модели сверточной нейронной сети и успешно обеспечивает защиту пользователей от спуфинг-атак.

Литература

1. Du Zhihao, Wang Yuxuan, Chen Qian, Shi Xian, Lv Xiang, Zhao Tianyu, Yang Zhifu Gao Yexin, Gao Changfeng, Wang Hui, Yu Fan, Liu Huadai, Gu Zhengyan Sheng Yue, Deng Chong, Wang Wen, Zhang Shiliang, Yan Zhijie, Zhou Jingren. CosyVoice 2: Scalable Streaming Speech Synthesis with Large Language Models // 2024. URL: arXiv:2412.10117v1

2. Поров А.В. Адаптивное скалярное квантование спектральных коэффициентов для системы сжатия аудио сигналов: автореферат на соискание ученой степени тех. наук: 05.13.01. - Санкт-Петербург, 2009. URL: new-disser.ru/_avtoreferats/01004571050.pdf

3. Хеин М.З. Современное состояние проблемы обработки, анализа и синтеза речевых сигналов // Computational nanotechnology. 2018. №2. URL: cyberleninka.ru/article/n/sovremennoe-sostoyanie-problemy-obrabotki-analiza-i-sinteza-rechevyh-signalov

4. Маршалко Д. А., Кубанских О. В. Архитектура свёрточных нейронных сетей // Ученые записки Брянского государственного университета. 2019. №4 (16). URL: cyberleninka.ru/article/n/arhitektura-svyortochnyh-neyronnyh-setey

5. Ранасингхе Н.К., Круглова Л.В. The Role of Convolutional Neural Networks in Cricket Performance Analysis // Вестник РУДН. Серия: Инженерные исследования. 2024. №2. URL: cyberleninka.ru/article/n/the-role-of-convolutional-neural-networks-in-cricket-performance-analysis

6. Пономарев, К.Г. Актуальные угрозы голосовых дипфейков для интернета вещей и методы их защиты // Исследовательский проект года 2024: сборник статей Международного научно-исследовательского конкурса, Петрозаводск, 02 декабря 2024 года. – Петрозаводск: Международный центр научного партнерства «Новая Наука» (ИП Ивановская И.И.), 2024. – С. 47-53. – EDN ETYDWJ.

7. Пономарев К.Г. Способы генерации голосовых дипфейков и методы их выявления // Молодежь. Наука. Инновации. – 2024. – Т. 1. – С. 172-176. – EDN QCFVLG.

8. Дюльдин Е.В., Зайцев К.С. Применение глубокого обучения для выявления и классификации DGA доменов // International Journal of Open Information Technologies. 2022. №8. URL: cyberleninka.ru/article/n/primenenie-glubokogo-obucheniya-dlya-vyyavleniya-i-klassifikatsii-dga-domenov

9. Стеценко А.И., Смагулова А.С. Разработка нейронной сети для распознавания изображений на базе Tensorflow // E-Scio. 2023. №3 (78). URL: cyberleninka.ru/article/n/razrabotka-neyronnoy-seti-dlya-raspoznavaniya-izobrazheniy-na-baze-tensorflow

10. Скрипачев В. О., Гуйда М. В., Гуйда Н. В., Жуков А. О. Особенности работы сверточных нейронных сетей // International Journal of Open Information Technologies. 2022. №12. URL: cyberleninka.ru/article/n/osobennosti-raboty-svertochnyh-neyronnyh-setey (дата обращения: 12.01.2025)

11. Пономарёв К.Г., Верещагина Е.А. Математический аппарат и технологическая инфраструктура системы прогнозирования синтетических голосовых дипфейков // Инженерный вестник Дона. - 2024. - №6. URL: ivdon.ru/ru/magazine/archive/n6y2024/9312

References

1. Du Zhihao, Wang Yuxuan, Chen Qian, Shi Xian, Lv Xiang, Zhao Tianyu, Yang Zhifu Gao Yexin, Gao Changfeng, Wang Hui, Yu Fan, Liu Huadai, Gu Zhengyan Sheng Yue, Deng Chong, Wang Wen, Zhang Shiliang, Yan Zhijie, Zhou Jingren. CosyVoice 2: Scalable Streaming Speech Synthesis with Large Language Models. 2024. URL: arXiv: 2412.10117v1

2. Porov A.V. Adaptivnoe skalyarnoe kvantovanie spektral`ny`x koe`fficientov dlya sistemy` szhatiya audio signalov [Adaptive scalar quantization of spectral

coefficients for audio signal compression system]: avtoreferat na soiskanie uchenoj stepeni tex. nauk: 05.13.01. Sankt-Peterburg, 2009. URL: new-disser.ru/_avtoreferats/01004571050.pdf

3. Xein M.Z. Computational nanotechnology. 2018. №2. URL: cyberleninka.ru/article/n/sovremennoe-sostoyanie-problemy-obrabotki-analiza-i-sinteza-rechevyh-signalov

4. Marshalko D. A., Kubanskix O. V. Ucheny`e zapiski Bryanskogo gosudarstvennogo universiteta. 2019. №4 (16). URL: cyberleninka.ru/article/n/arhitektura-svyortochnyh-neyronnyh-setey

5. Ranasingxe N.K., Kruglova L.V. Vestnik RUDN. Seriya: Inzhenerny`e issledovaniya. 2024. №2. URL: cyberleninka.ru/article/n/the-role-of-convolutional-neural-networks-in-cricket-performance-analysis

6. Ponomarev, K.G. Issledovatel`skij proekt goda 2024: sbornik statej Mezhdunarodnogo nauchno-issledovatel`skogo konkursa, Petrozavodsk, 02 dekabrya 2024 goda. Petrozavodsk: Mezhdunarodny`j centr nauchnogo partnerstva «Novaya Nauka» (IP Ivanovskaya I.I.), 2024. pp. 47-53. pEDN ETYDWJ.

7. Ponomarev K.G. p Molodezh`. Nauka. Innovacii. p. 2024. T. 1. pp. 172-176. EDN QCFVLG.

8. Dyul`din E.V., Zajcev K.S. International Journal of Open Information Technologies. 2022. №8. URL: cyberleninka.ru/article/n/primenenie-glubokogo-obucheniya-dlya-vyyavleniya-i-klassifikatsii-dga-domenov

9. Stecenko A.I., Smagulova A.S. E-Scio. 2023. №3 (78). URL: cyberleninka.ru/article/n/razrabotka-neyronnoy-seti-dlya-raspoznavaniya-izobrazheniy-na-baze-tensorflow

10. Skripachev V. O., Gujda M. V., Gujda N. V., Zhukov A. O. International Journal of Open Information Technologies. 2022. №12. URL: cyberleninka.ru/article/n/osobennosti-raboty-svertochnyh-neyronnyh-setey (data obrashheniya: 12.01.2025).

11. Ponomaryov K.G., Vereshhagina E.A. Inzhenernyj vestnik Dona. 2024. №6.

Дата поступления: 13.12.2024

Дата публикации: 2.02.2025