МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«Южно-Уральский государственный университет
(национальный исследовательский университет)»
Высшая школа электроники и компьютерных наук
Кафедра системного программирования

ОТЧЕТ
по производственной практике
(научно-исследовательская работа)

Научный руководитель,
Доцент кафедры СП,к.ф.-м.н.
_____ Г. И. Радченко

Автор работы,
студент группы КЭ-403
_____ В. А. Дегтярев

Челябинск, 2023 г.

Министерство науки и высшего образования Российской Федерации
Южно-Уральский государственный университет
Кафедра системного программирования

УТВЕРЖДАЮ

Зав. кафедрой
системного программирования

_____ Л.Б. Соколинский

ЗАДАНИЕ
на производственную практику
(научно-исследовательскую работу)

1. Тема работы

Проектированние методов машинного анализа географически-распределенных данных

2. Исходные данные к работе

2.1. PySyft documentation [Электронный ресурс] URL: https://openmined.github.io/PySyft

3. Перечень подлежащих разработке вопросов

3.1. Выполнить обзор литературы;
3.2. Выполнить обзор существующих аналогов;
3.3. Спроектировать методы машинного анализа географически-распределенных данных

4. Сроки

Дата выдачи задания: 1 февраля 2023 г.
Срок сдачи законченной работы: 20 февраля 2023 г.

Руководитель практики со стороны ЮУрГУ:

Доцент кафедры СП,к.ф.-м.н. _____ Турлакова С.У.
подпись ФИО ответственного

Научный руководитель практики:

Доцент кафедры СП,к.ф.-м.н. _____ Радченко Г. И.
должность, ученая степень подпись ФИО научного руководителя

Задание принял к исполнению:

_____ Дегтярев В.А.
подпись ФИО студента

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ 4
1. АНАЛИЗ ПРЕДМЕТНОЙ ОБЛАСТИ 6
1.1. Описание предметной области 6
1.2. Анализ аналогичных проектов 8
1.2.1. PySyft 8
1.2.2. TensorFlow Federated 9
2. ПРОЕКТИРОВАНИЕ 11
2.1. Функциональные требования к системе 11
2.2. Нефункциональные требования к системе 12
2.3. Диаграмма вариантов использования системы 12
2.4. Диаграмма компонентов 14
ЗАКЛЮЧЕНИЕ 16
ЛИТЕРАТУРА 17

ВВЕДЕНИЕ

Актуальность

В настоящее время в мире существует огромное количество информации. А также большое количество людей, которые хотят использовать эту информацию в своих научных интересах, для нахождения закономерностей, обучения нейросетей, чтобы получать ответы на важные вопросы.

Машинное обучение и анализ данных уже активно применяется в медицине, финансах, промышленности [1]. Однако эти технологии еще не могут уверенно отвечать на некоторые глобальные и сложные вопросы из-за отсутствия доступа у разработчиков и ученых к большому количеству информации. Основные причины этого выражаются в виде защиты персональных данных, сохранения приватности конкретных данных, а также раздробленности этих данных среди огромного количества организаций.

Эти проблемы можно решить с помощью систем географически-распределенного и конфиденциального машинного обучения.

Постановка задачи

Целью выпускной квалификационной работы является реализация системы на основе методов машинного анализа географически-распределенных данных. Для достижения поставленной цели необходимо решить следующие задачи:

1) выполнить обзор литературы;
2) выполнить анализ аналогичных проектов;
3) определить функциональные и нефункциональные требования к системе;
4) спроектировать методы машинного анализа географически-распределенных данных;

5) реализовать методы машинного анализа географически-распределенных данных;

6) провести тестирование методов машинного анализа географически-распределенных данных.

Структура и содержание работы

Работа состоит из введения, пяти глав, заключения и списка литературы. Объем работы составляет 29 страниц, объем списка литературы – 11 источников.

В первой главе описываются предметная область и аналогичные проекты.

Вторая глава содержит описание теоретической части по теме работы.

Третья глава посвящена определению функциональных и нефункциональных требований к системе и проектированию ее архитектуры.

Четвертая глава содержит в себе подробности и особенности реализации методов машинного анализа.

В пятой главе описывается процесс тестирования работы методов машинного анализа географически-распределенных данных.

1. АНАЛИЗ ПРЕДМЕТНОЙ ОБЛАСТИ

1.1. Описание предметной области

Целью данной работы является разработка системы на основе методов машинного анализа географически-распределенных данных. Анализ данных представляет собой область математики и информатики, которая занимается построением и исследованием наиболее общих математических методов и вычислительных алгоритмов извлечения знаний из экспериментальных данных с целью получения полезной информации и принятия решений [2].

Для решения сложных аналитических задач часто используются нейронные сети. Нейронная сеть – это математическая модель, а также её программное воплощение, построенная по принципу организации и функционирования биологических нейронных сетей [3]. Нейронная сеть обучается в процессе обработки входных данных и при постепенном подборе нужных коэффициентов. Моделью обучения называется файл, который обучен распознаванию определенных типов закономерностей. Модель обучается на основе набора данных и алгоритма, который она может использовать для анализа и обучения на основе этих данных [4]. Нейронная сеть не имеет последовательного алгоритма выполнения. Результаты анализа нейронной сетью идентичных данных могут различаться между собой.

Особенностью машинного анализа географически-распределенных данных является то, что информация, которую анализирует система, распределена между различными независимыми устройствами.

Для обеспечения географически-распределенного анализа и конфиденциальности данных современные решения используют такие методы, как федеративное обучение, дифференциальная приватность, гомоморфное шифрование [5].

Федеративное обучение представляет собой метод машинного обучения, который позволяет коллективно обучать алгоритм на нескольких устройствах без централизации всех исходных данных на одном сервере. Системы федеративного обучения совершенствуют единую общую модель. Источники данных никогда не перемещаются и не объединяются, но каждое устройство вносит свой вклад в обучение и повышение качества общей модели [6].

Дифференциальная приватность – это область, изучающая методы, которые обеспечивают максимально точные результаты статистических запросов в базу данных при минимизации возможности идентификации отдельных записей в ней. Для каждого человека, чьи данные входят в анализируемый набор, дифференциальная приватность гарантирует, что результат анализа на дифференциальную приватность будет практически неотличим вне зависимости от того, есть ли данные этого конкретного человека в наборе или нет [7]. Дифференциальная приватность основана на введении случайности в данные. Чем больше случайности добавляется, тем сильнее сохраняется приватность, однако получаются более неточные результаты. Также на точность результатов влияет размер выборки, чем больше выборка, тем точнее результаты [8].

Гомоморфное шифрование – это форма шифрования, позволяющая производить определённые математические действия с зашифрованным текстом и получать зашифрованный результат, который соответствует результату операций, выполненных с открытым текстом [9]. Использование гомоморфного шифрования открывает множество перспектив при обработке конфиденциальных данных в среде, участники которой не доверяют друг другу. Оно позволяет осуществлять индексацию, фильтрацию спама, обработку платежей и другие действия без расшифровки самих сообщений и может применятся в облачных вычислениях, децентрализованных системах, электронном голосовании [10].

1.2. Анализ аналогичных проектов

1.2.1. PySyft

В 2019 году была создана библиотека PySyft сообществом OpenMined. Это люди, объединенные темой конфиденциальности в машинном обучении. PySyft представляет собой обертку над PyTorch, Tensorflow или Keras для приватного машинного обучения [8].

Основная задача, стоящая перед сообществом OpenMined, заключалась в том, чтобы создать программное обеспечение, которое бы позволяла одному человеку получать ответы на свои вопросы, используя данные, принадлежащие другому человеку без необходимости просмотра и создания копии этих данных [11].

Проект предоставляет удаленный вызов процедур, что позволяет разработчику отправлять модель нейронной сети к пользователям, где она локально обучается на их данных, после чего возвращается с обновленными весами обратно разработчику (рисунок 1). Данный процесс может происходить одновременно на разных устройствах, тем самым происходит параллельное обучение модели нейронной сети[8].

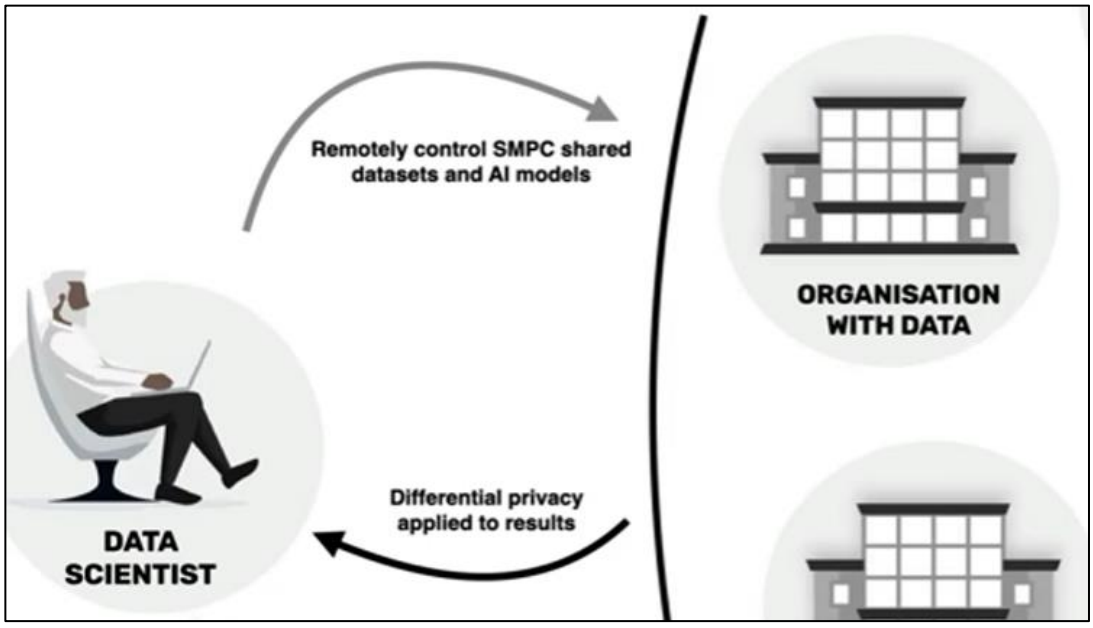

Рисунок 1 – Взаимодействие разработчика с пользователями [12]

Также ключевой особенностью PySyft является использование дифференциальной приватности [11]. По измененным весам модели нейронной сети можно догадаться, какие данные были у пользователя. Чтобы это предотвратить, к данным, которые хранятся на вычислительной машине пользователя, добавляется шум. Дифференциальная приватность представляет собой методы, которые описывают добавление шума.

1.2.2. TensorFlow Federated

Компания Google давно занимается сбором некоторой информации с устройств пользователей в единое защищенное хранилище, на котором тренируют свои нейросети. А в 2017 году ученые из Google Research предложили инновационный подход под названием федеративное машинное обучение. Он позволяет всем устройствам, которые участвуют в машинном обучении, делить на всех единую модель для прогнозирования, но при этом не делиться первичными данными для обучения модели. Система федеративного обучения работает по принципу совершенствования единой общей модели нейросети [13].

Для проверки системы федеративного обучения на больших объемах данных компанией Google был реализовал этот алгоритм в мобильном приложении клавиатуры Gboard. Основной задачей являлось прогнозирование слов и выражений, которые пользователь предположительно использовал бы следующими во время печатания текста. Система федеративного обучения не отправляет текст, который печатает пользователь, на сервер компании Google, она отправляет На устройстве каждого пользователя производится анализ текста, который он использовал. Затем результаты анализа отправляются в компанию Google, где они будут объединены с другими результатами анализа для улучшения общей модели

набора текста. Тем самым, каждый пользователь улучшает опыт использования Gboard каждому пользователю.

Компанией Google была создана платформа TensorFlow Federated с открытым исходным кодом для машинного обучения на децентрализованных данных. Архитектура данного платформы представлена на рисунке 2. TensorFlow Federated был разработан для облегчения исследований и экспериментов с федеративным обучением [14].

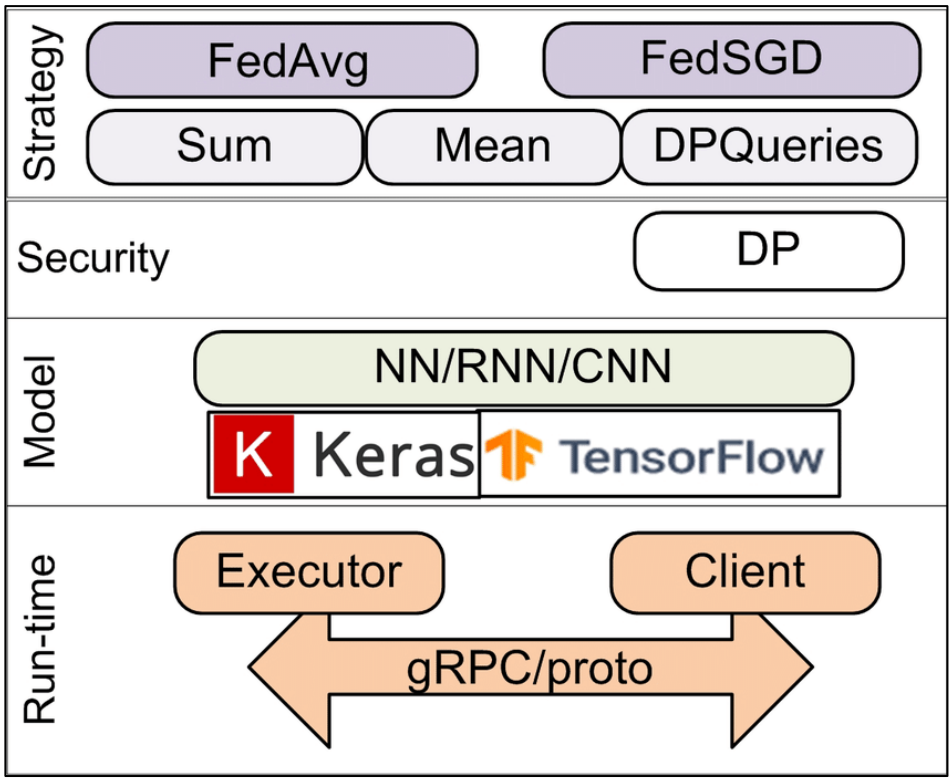

Рисунок 2 – Архитектура TensorFlow Federated [14]

Выводы по первой главе

В данной главе был проведен обзор предметной области и анализ существующих решений и проектов в области машинного анализа географически-распределенных данных.

2. ПРОЕКТИРОВАНИЕ

Целью данной работы является разработка системы для анализа географически-распределенных данных на платформе PySyft. Система представляет собой децентрализованное приложение, которое будет предоставлять одноранговую сеть для владельцев данных и аналитиков данных.

Владелец данных с помощью приложения сможет создать узел внутри целостной одноранговой сети, загружать и управлять данными, расположенными на этом узле.

Аналитик данных с помощью приложения сможет подключаться к различным узлам внутри сети и проводить аналитические операции на основе данных, расположенным на этих узлах.

В рамках данной работы будет создано 6 узлов с данными. Также будет создан 1 узел для анализа данных, которые расположены на других узлах.

2.1. Функциональные требования к системе

Можно выделить следующий набор функциональных требований к системе.

1. Система должна предоставлять владельцу данных возможность запустить узел в одноранговой сети.
2. Система должна предоставлять владельцу данных возможность загрузить данные на созданный им узел в одноранговой сети.
3. Система должна предоставлять владельцу данных возможность удалить данные на созданном им узле в одноранговой сети.
4. Система должна предоставлять владельцу данных возможность отключить созданный им узел в одноранговой сети.

5. Система должна предоставлять владельцу данных возможность указать, к каким данные, среди тех, которые владелец данных разместил на узле, ограничивать доступ для аналитика данных.

6. Система должна предоставлять аналитику данных возможность использовать данные, размещенные на любых других узлах одноранговой сети.

2.2. Нефункциональные требования к системе

Можно выделить следующие нефункциональные требования к системе.

1. Система должна образовывать одну целостную одноранговую сеть.

2. Система должна обеспечивать аналитику данных возможность получать данные из нескольких узлов одновременно.

3. Система должна быть написана на языке программирования Python.

4. Система должна быть разработана с использованием таких инструментов, как: PySyft, PyGrid, PyTorch.

2.3. Диаграмма вариантов использования системы

Для проектирования системы был использован язык графического описания для объектного моделирования UML. На рисунке 3 представлена диаграмма вариантов использования.

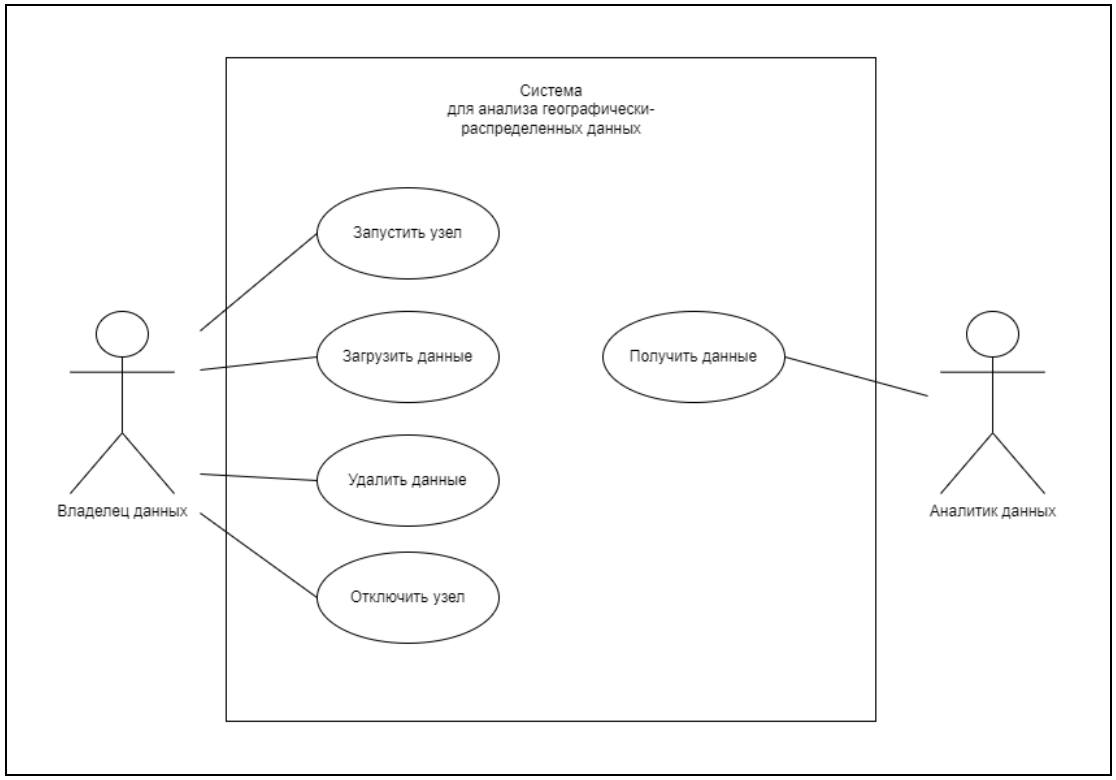

Рисунок 3 – Диаграмма вариантов использования системы для анализа географически-распределенных данных

В системе определены следующие виды акторов.

1. *Владелец данных* – это пользователь приложения, который может запустить и отключить узел в одноранговой сети для размещения на него определенных данных.

2. *Аналитик данных* – пользователь приложения, который может использовать размещенные на узлах данные для аналитических операций.

Актору «Владелец данных» доступны следующие варианты использования системы.

1. Владелец данных может запустить узел в одноранговой сети с помощью приложения для дальнейшего размещения на нем данных.

2. Владелец данных может загрузить данные на созданный им узел, которые в дальнейшем будут доступны аналитикам данных.

3. Владелец данных может удалить данные с созданного им узла.

4. Владелец данных может отключить узел в одноранговой сети с помощью приложения, после этого, все данные, которые были размещены на узле, станут недоступными для использования аналитиками данных.

Актор «Аналитик данных» может использовать систему только для получения данных, которые разместил актор «Владелец данных». Для этого ему необходимо выбрать определенный узел, загрузить данные, которые на нем размещены, и использовать их в своих аналитических операциях.

2.4. Диаграмма компонентов

На рисунке 4 представлена диаграмма компонентов системы.

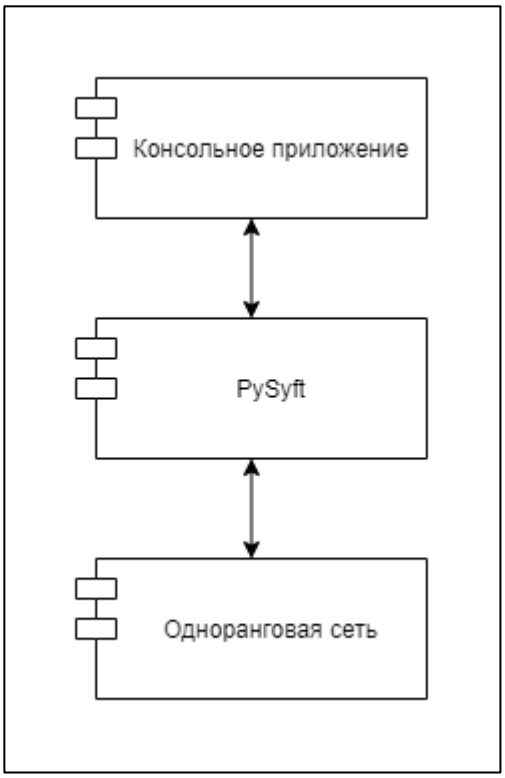

Рисунок 6 – Диаграмма компонентов системы

Система состоит из следующих компонентов.

1. Консольное приложение – приложение, с которым взаимодействуют пользователи.

2. PySyft – платформа, которая предоставляет методы для безопасного и конфиденциального анализа данных.

3. Одноранговая сеть – сетевая технология, которая позволяет нескольким сетевым устройствам совместно использовать ресурсы и взаимодействовать друг с другом.

ЗАКЛЮЧЕНИЕ

В рамках данной работы была спроектирована система для анализа географически-распределенных данных на платформе PySyft. При этом были решены следующие задачи.

1. Выполнен обзор литературы.
2. Выполнен анализ аналогичных проектов.
3. Определены функциональные и нефункциональные требования к системе.

ЛИТЕРАТУРА

1. 7 примеров применения машинного обучения в 5 отраслях бизнеса. [Электронный ресурс] URL: https://mcs.mail.ru/blog/17-primerov-mashinnogo-obucheniya (дата обращения: 13.02.2023 г.).

2. Анализ данных – основы и терминология. [Электронный ресурс] URL: https://habr.com/ru/post/352812/ (дата обращения: 13.02.2023 г.).

3. Нейронные сети для начинающих. Решение задачи классификации Ирисов Фишера. [Электронный ресурс] URL: https://habr.com/ru/company/ruvds/blog/679988/ (дата обращения: 16.02.2023 г.).

4. Что такое модель машинного обучения? [Электронный ресурс] URL: https://learn.microsoft.com/ru-ru/windows/ai/windows-ml/what-is-a-machine-learning-model (дата обращения: 16.02.2023 г.).

5. Официальный сайт PySyft. [Электронный ресурс] URL: https://github.com/OpenMined/PySyft/ (дата обращения: 13.02.2023 г.).

6. Масштабируемый подход к частично локальному федеративному обучению. [Электронный ресурс] URL: https://habr.com/ru/post/645783/ (дата обращения: 16.02.2023 г.).

7. Дифференциальная приватность — анализ данных с сохранением конфиденциальности. [Электронный ресурс] URL: https://habr.com/ru/company/domclick/blog/526724/ (дата обращения: 16.02.2023 г.).

8. Конфиденциальное машинное обучение. Библиотека PySyft. [Электронный ресурс] URL: https://habr.com/ru/post/500154/ (дата обращения: 16.02.2023 г.).

9. Методы обфускации трафика. Гомоморфное шифрование. [Электронный ресурс] URL: https://habr.com/ru/company/globalsign/blog/717482/ (дата обращения: 16.02.2023 г.).

10. Цифровые фиатные деньги, гомоморфное шифрование и другие перспективные направления криптографии. [Электронный ресурс] URL: https://habr.com/ru/company/kryptonite/blog/658113/ (дата обращения: 16.02.2023 г.).

11. Официальный сайт OpenMined. [Электронный ресурс] URL: https://www.openmined.org/ (дата обращения: 13.02.2023 г.).

12. Introduction to Remote Data Science. [Электронный ресурс] URL: https://habr.com/ru/post/402987/ (дата обращения: 13.02.2023 г.).

13. Google изобрела распределённый ИИ для миллиарда смартфонов. [Электронный ресурс] URL: https://habr.com/ru/post/402987/ (дата обращения: 13.02.2023 г.).

14. Официальный сайт Tensorflow Federated. [Электронный ресурс] URL: https://www.tensorflow.org/federated/ (дата обращения: 13.02.2023 г.).