|  |  |
| --- | --- |
|  | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ \_\_ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА \_\_\_\_СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ\_\_\_\_\_\_\_\_\_\_\_\_

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_*** ***Классификация данных на спутниковых\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_снимках \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

Студент \_\_\_\_ИУ5-63Б\_\_\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_Требуков\_Д.А.\_\_\_**

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

Консультант **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

*2023 г.*

**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение**

**высшего образования**

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

**(национальный исследовательский университет)»**

**(МГТУ им. Н.Э. Баумана)**

УТВЕРЖДАЮ

Заведующий кафедрой \_\_\_\_\_\_\_\_\_\_

(Индекс)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(И.О.Фамилия)

« \_\_\_\_\_ » \_\_\_\_\_\_\_\_\_\_\_\_ 20 \_\_\_\_ г.

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

по теме \_\_\_\_\_\_\_\_\_\_\_Классификация данных на спутниковых снимках\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент группы \_\_\_\_ИУ5-63Б\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Требуков Даниил Андреевич\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Источник тематики (кафедра, предприятие, НИР) \_\_\_\_\_\_\_кафедра\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

График выполнения НИР: 25% к \_\_\_ нед., 50% к \_\_\_ нед., 75% к \_\_ нед., 100% к \_\_\_ нед.

Техническое задание \_Исследовать методы классификации данных с помощью машинного обучения для выявления повреждений на спутниковых снимках по заданному классу \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Оформление научно-исследовательской работы:***

Расчетно-пояснительная записка на \_\_22\_\_ листах формата А4.

Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата выдачи задания « \_\_\_ » \_\_\_\_\_\_\_\_\_\_\_\_ 20\_\_ г.

**Руководитель НИР**  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_Требуков Д.А.\_\_\_**

(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

**СОДЕРЖАНИЕ**

[1. Постановка задачи 5](#_Toc136978259)

[2. Исследование данных 7](#_Toc136978260)

[2.1. Описание и структура данных 7](#_Toc136978261)

[2.2. Подготовка данных 10](#_Toc136978262)

[2.3. Визуализация данных 11](#_Toc136978263)

[3. Описание метода обучения 15](#_Toc136978264)

[3.1. Применение алгоритма RandomForest 15](#_Toc136978265)

[3.2. Оценка обученной модели 16](#_Toc136978266)

[3.3. Фильтр Габора 17](#_Toc136978267)

[Заключение 20](#_Toc136978268)

[Список литературы 21](#_Toc136978269)

**ВВЕДЕНИЕ**

В последнее время все большую актуальность приобретают методы зондирования крупных лесных массивов с помощью орбитальных спутников. Информацию о земной поверхности используют для мониторинга и исследования состояния окружающей среды. При этом с помощью машинного обучения, становиться возможным оперативный контроль за состоянием окружающей среды. Например, автоматическое выявление лесных пожаров.

В современных исследованиях по оценке и прогнозированию состояния окружающей среды территорий активно используются мультиспектральные изображения со спутников Sentinel-2, которые ведут съёмку не только в 3 привычных каналах видимого спектра (синий, зеленый и красный), но также и в невидимых глазу диапазонах: коротковолновом инфракрасном (SWIR), видимом ближнем инфракрасном (VNIR) и т.д. Благодаря тому, что спутники Sentinel-2 имеют самое высокое разрешение из свободно распространяемых – 10 метров, они могут различать разные типы поверхности (вода, лес, дороги, здания) и отслеживать лесные пожары, вырубки.

С учетом природных особенностей ландшафта и текстурных характеристик территорий мониторинга, для спутниковых снимков применяются классические методы обучения, позволяющие учитывать эти особенности. Для этого у каждого космоснимка сформированы маски в формате *geoTiff*. GeoTiff является подмножеством формата хранения растровых графических изображений tiff, в котором содержится дополнительная геоинформация.

Целью данной научно-исследовательской работы является создание модели автоматического определения лесных дорог на изображениях со спутника Sentinel-2.

Для достижения цели необходимо выполнить следующие задачи:

1. Исследовать и отобрать полученные данные
2. Применить методы создания маски изменения снимков
3. Подготовить данные (удалить из снимков классы, не относящиеся к варианту, использовать NDVI и определить обучающие и тестовые выборки)
4. Визуализировать данные
5. Обучить модель с помощью алгоритма RandomForest
6. Улучшить результаты обученной модели с помощью фильтра Габора
7. **ПОСТАНОВКА ЗАДАЧИ**

Задача данной работы заключается в обучении модели бинарной классификации, для выявления повреждений на основе сравнения старых и новых снимков со спутника Sentinel-2. В выбранном варианте классом бинарной сегментации станут лесные дороги (класс 3).

Sentinel-2 – это пара сверхразрешающих многозадачных спутников, которые были разработаны Европейским космическим агентством для обеспечения непрерывного земного наблюдения и мониторинга природных ресурсов [1]. Данные снимки, полученные со спутника Sentinel-2, используются для многих приложений, включая земледелие, экологию, лесоводство, геологию, картографию и планирование городской застройки. Анализ космических снимков со спутника Sentinel-2 включает в себя множество методов обработки данных, таких как классификация изображений, мозаики, обнаружение изменений на земле, мониторинг и диагностику взаимодействия между биозонами [2].

Один из примеров использования данных от Sentinel-2 – это анализ урожайности полей. По снимкам получают информацию о том, сколько участков земли было посеяно и вырастило хороший урожай, а сколько пострадало от засухи, ветрового надвига песка или насекомых-вредителей. Это позволяет фермерам быстро принимать решение о том, нуждаются ли поля в дополнительном поливе или обработке, чтобы сохранить урожай [3].

Еще один пример использования космических снимков со спутника Sentinel-2 – это мониторинг растительного покрова и изменений экосистем. По снимкам можно видеть, как меняется зеленая масса за сезон или какие участки леса были вырублены [4]. Эти данные могут быть полезными для различных исследований, в том числе для оценки состояния окружающей среды и слежения за глобальными климатическими изменениями.

Кроме того, космические снимки со спутника Sentinel-2 могут использоваться для планирования городской застройки и развития инфраструктуры. По снимкам можно узнать, какое количество земли занято жилыми домами, сколько производственных площадок существует и как распределено зеленое насаждение в городе. Основываясь на этих данных, можно разрабатывать новые проекты, учитывая специфики местности.

В целом, космические снимки со спутника Sentinel-2 предоставляют богатый и точный инструмент для мониторинга и планирования землепользования, оценки состояния окружающей среды и исследования глобальных изменений климата.

Ранее уже проводились исследования по мониторингу лесов на основе снимков спутника Sentinel-2, предоставляющего данные оптических изображений высокого разрешения для поддержки исследований компаний по отслеживанию состояния земли, включая мониторинг растительности, почвы и водного покрова, а также наблюдение за внутренними водными путями и прибрежными районами. В одном из исследований Центра устойчивого управления и дистанционного мониторинга лесов Поволжского государственного технологического университета было произведено распознавание лесных насаждений и доминирующих древесных пород на территории Юрсовского и Ленинского лесничеств Пензенской области и предоставлены данных для картографирования (дешифрирования) породного состава лесных насаждений на основе полученных космоснимков [5].

В данной научно-исследовательской работе для выявления повреждений леса будут использоваться многоспектральные снимки спутника Sentiel-2, разрезанные на части 256×256 пикселей и объединённые в пары старый-новый. Для исследования будут составлены маски изменения области леса на основе хранимой внутри снимков геоинформации. Для оценки растительности будет применён индекс NDVI, так как в исследовании есть снимки с вырубкой леса и другими повреждениями. Для классификации снимков будут использованы алгоритмы RandomForest и XGBoost.

В рамках исследования будет использован Jupyter Notebook, а также такие библиотеки, как numpy, pandas, scilit-learn, matplotlib, skimage, tqdm и cv2. Они необходимы для преобразования снимков в нужный формат, открытия файлов, визуализации данных и пр. Методы машинного обучения, такие как RandomForest, необходимые в исследовании, уже реализованы в библиотеке scilit-learn и будут использованы для классификации повреждений леса.

В научной работе будет проведена классификация областей леса на повреждённые и неповреждённые, то есть бинарная классификация. Таким образом, исследование будет требовать обработку снимков (удаление неиспользуемых классов из геоинформации снимков, добавление каналов). Также для оценки обученной модели будет использоваться метрика Dice.

1. **ИССЛЕДОВАНИЕ ДАННЫХ**

В исследовательской части будет производиться изучение данных со спутниковых снимков, включая: описание полученных данных, подготовка данных и их визуализация. Изначально данные находятся в архиве в виде изображений со спутника Sentinel-2 в формате geoTiff, представляющие собой растровые графические изображения tiff, с содержащейся в них дополнительной геоинформацией. Для получения этой геоинформации (например, координат снимков) можно использовать множество библиотек, в том числе GDAL, но для текущих задач это не нужно.

* 1. **ОПИСАНИЕ И СТРУКТУРА ДАННЫХ**

Данные представляют из себя многоспектральные спутниковые снимки, поделённые на части по 256×256 пикселей и объединённые в пары старый-новый. Так как изображения представляют из себя матрицу (состоящую из пикселей), для обнаружения лесных повреждений достаточно сравнить старый и новый снимок по пикселям. В простейшем случае для этого достаточно вычесть одно изображение из другого (это же матрицы), а затем рассмотреть, для каких пикселей разность сильно отличается от нуля. Однако нужно учесть, что в реальности дело осложняется разными условиями съёмки, дефектами и т.д., так что не лишним является применение методов машинного обучения.

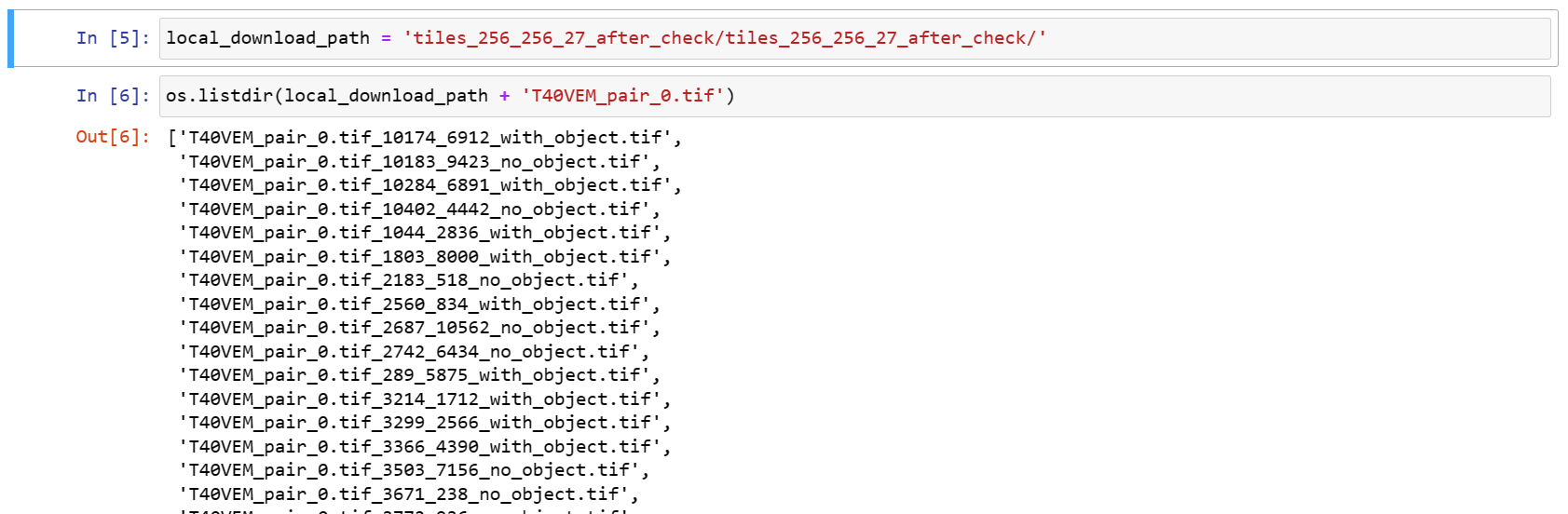
Все изображения распределены в папках “tile\_256\_256\_27\_add”, “tile\_256\_256\_27\_additionegor”, “tile\_256\_256\_27\_additionegor\_noobjmask”, “tile\_256\_256\_27\_after\_check”, “tile\_256\_256\_27\_windthrows”. В каждой из этих папок находятся папки для больших снимков, разделённых на части. Для примера рассмотрим содержание папки “tile\_256\_256\_27\_after\_check” (рис. 2.1).

Рисунок 2.1 – Содержимое папки “tile\_256\_256\_27\_after\_check”

Для удобства при дальнейшей работе объединим пути ко всем изображениям в один список, выбрав только те пары снимков, где содержаться объекты для исследования. Интересующие нас файлы содержат «with\_object» в названии (рис. 2.2).

Рисунок 2.2 – Список данных

Посмотрим на размерность одного из изображений (рис. 2.3).

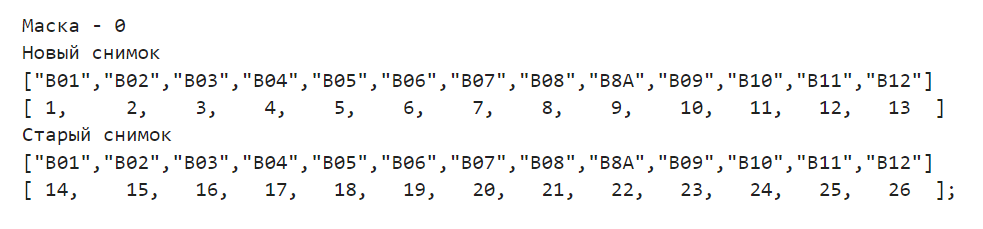
Рисунок 2.3 – Размерность изображений

Видим, что изображение – это трёхмерный массив. 256 пикселей по вертикали и по горизонтали, а также 27 каналов "в глубину". Sentinel-2 ведет съёмку не только в 3 привычных каналах видимого спектра (синий, зеленый и красный), но также и в невидимых глазу диапазонах: коротковолновом инфракрасном (SWIR), видимом ближнем инфракрасном (VNIR) и т.д., - всего 13 каналов (таблица 1).

Таблица 1 – Каналы невидимого спектра излучения

|  |  |  |  |
| --- | --- | --- | --- |
| Канал | Разрешение | Центральная длина волны | Описание |
| B1 | 60 m | 443 nm | Ultra Blue (Coastal and Aerosol) |
| B2 | 10 m | 490 nm | Blue |
| B3 | 10 m | 560 nm | Green |
| B4 | 10 m | 665 nm | Red |
| B5 | 20 m | 705 nm | Visible and Near Infrared (VNIR) |
| B6 | 20 m | 740 nm | Visible and Near Infrared (VNIR) |
| B7 | 20 m | 783 nm | Visible and Near Infrared (VNIR) |
| B8 | 10 m | 842 nm | Visible and Near Infrared (VNIR) |
| B8a | 20 m | 865 nm | Visible and Near Infrared (VNIR) |
| B9 | 60 m | 940 nm | Short Wave Infrared (SWIR) |
| B10 | 60 m | 1375 nm | Short Wave Infrared (SWIR) |
| B11 | 20 m | 1610 nm | Short Wave Infrared (SWIR) |
| B12 | 20 m | 2190 nm | Short Wave Infrared (SWIR) |

В наших данных заранее совмещены пары старых снимков (со все ещё с нетронутым лесом) и новых (с повреждением), а также маска (с отмеченным повреждением). Так что всего получается 13+13+1=27 каналов. Порядок их следования в нашем трехмерном массиве изображен на рисунке 2.4.

Рисунок 2.4 – Каналы изображений

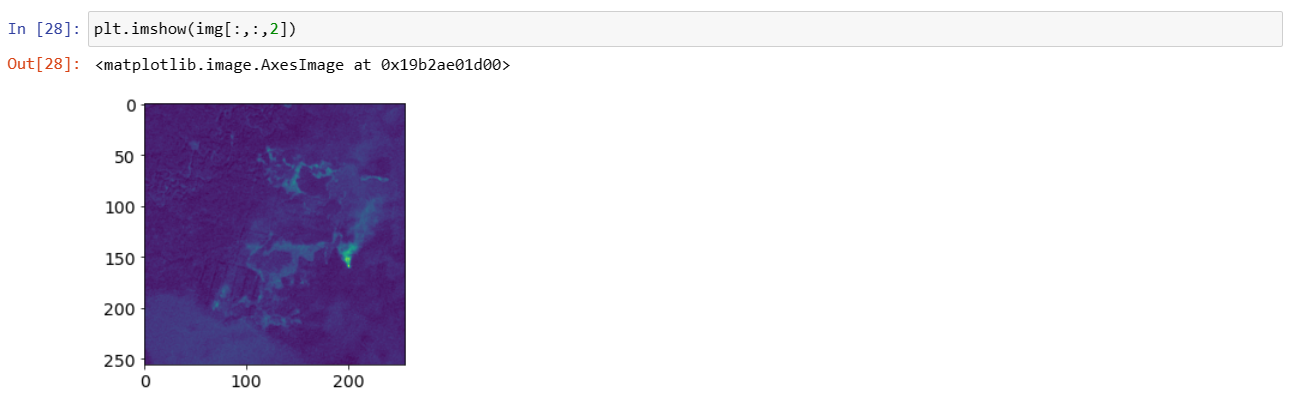
В маске разными цифрами указаны следующие классы:

* 1 - Сплошная рубка
* 2 - Проходная рубка
* 3 - Лесная дорога
* 4 - Ветровал
* 5 - Пожар
* 6 - Усыхание (повреждение)
* 7 - Выборочная рубка

Мы собираемся обучать модели всего лишь для бинарной классификации, так что достаточно будет проверять, больше ли нуля значение маски. Посмотрим количество *тайлов* (изображений с ландшафтом леса) по каждому классу в полном наборе данных (рис. 2.5).

Рисунок 2.5 – Количество снимков по классам

Видно, что для интересующего нас варианта с лесными дорогами (класс 3) было найдено 343 пары снимков. Попробуем посмотреть на один из каналов снимка с лесной дорогой. Например, выведем изображение через 2 канал, соответствующий синему спектру нового снимка (рис. 2.6).

Рисунок 2.6 – Изображение снимка через 2 канал

* 1. **ПОДГОТОВКА ДАННЫХ**

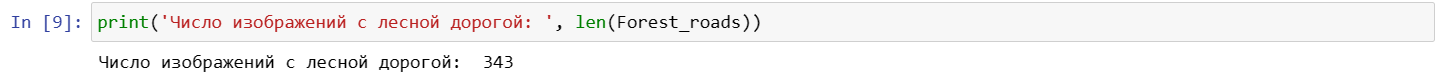
Для выбранного варианта с лесными дорогами мы будем обучать модель искать повреждения по классу 3, для этого соберем все пары снимков с этим классом в один массив *Forest\_roads* (рис.2.5). Количество собранных пар снимков изображено на рисунке 2.7.

Рисунок 2.7 – Количество изображений с лесной дорогой

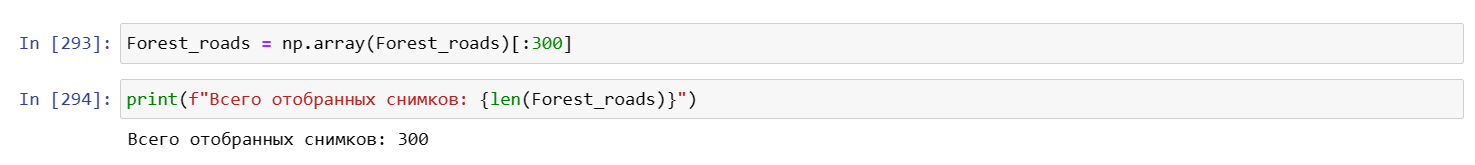
Для работы возьмем 300 изображений, этого количества достаточно для обучения модели. Отбирать снимки будем из созданного массива *Forest\_roads*, так как класс лесных дорог сильно отличается от остальных, и модель не будет их путать с другими классами (рис. 2.8).

Рисунок 2.8 – Отбор снимков для обучения модели

Сразу же отделим несколько директорий целиком для тестирования модели в дальнейшем. В каждой директории содержится один снимок, разрезанный на части, и если мы разделим данные случайно, части одного снимка могут оказаться и в тренировочных, и в тестовых данных - тогда у модели будет косвенная информация о тесте (хотя бы об общей яркости-контрастности изображений), что может завысить метрики (рис. 2.9).

Рисунок 2.9 – Создание тренировочной и тестовой выборок

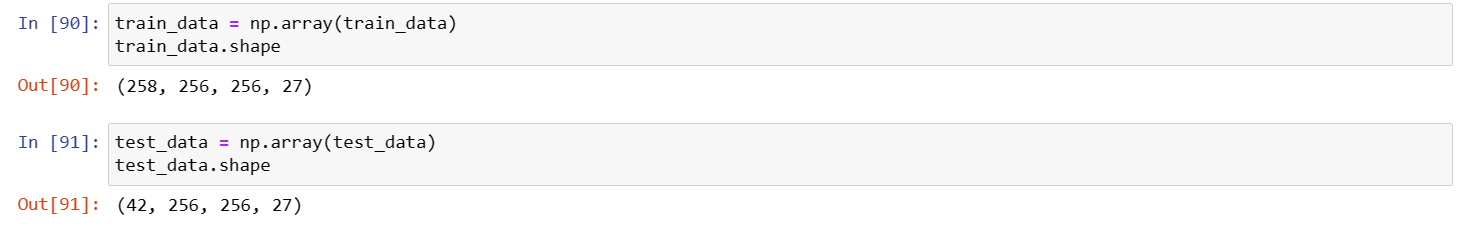
 Выбранные датасеты для обучения и тестирования наших будущих моделей изображены на рисунке 2.10.

Рисунок 2.10 – Данные для тестирования модели

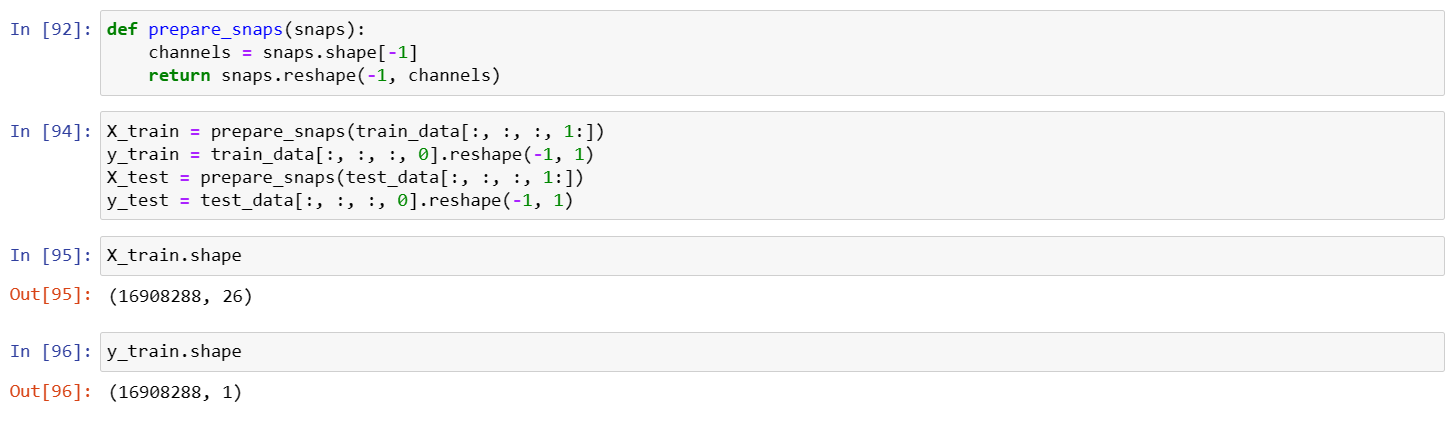
 На вход классических МЛ-моделей должен подаваться одномерный массив. Так что нам придётся "вытянуть" наши изображения 256х256х27 с помощью функции *reshape* из библиотеки NumPy, которая используется для изменения формы любого существующего массива без изменения данных, находящихся в нем [6]. Посмотрим на полученные обучающую и тестовую выборки (рис. 2.11).

Рисунок 2.11 – Обучающая и тестовая выборки

* 1. **ВИЗУАЛИЗАЦИЯ ДАННЫХ**

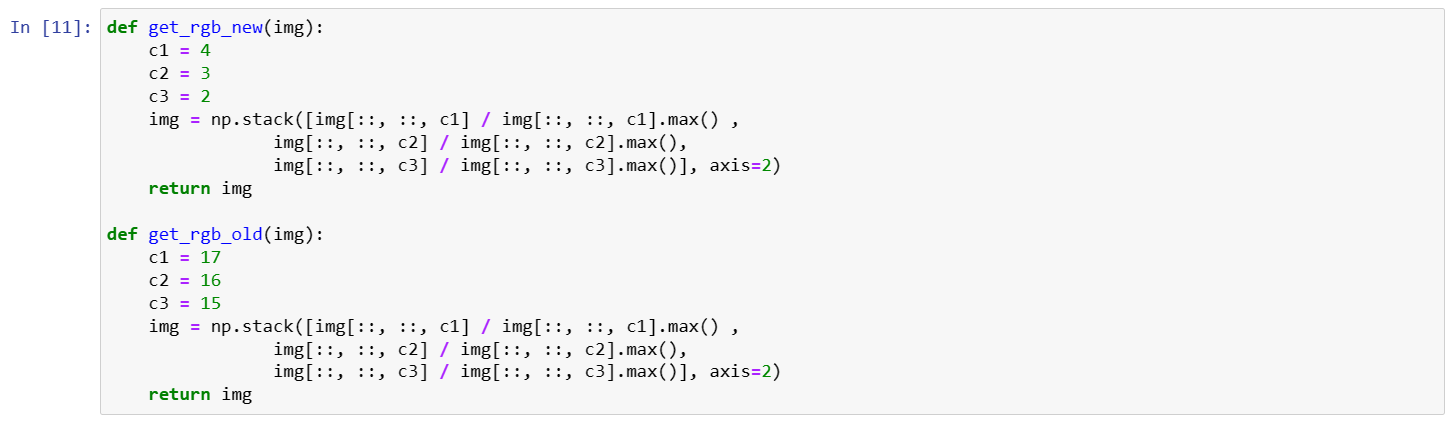
Ранее мы уже выводили изображение одного из полученных снимков леса через его второй канал, соответствующий синему спектру, от чего полученное изображение было преимущественно в синем цвете. Однако привычнее было бы смотреть на RGB-изображение, так что создадим функции для выделения старого и нового набора rgb-каналов (рис. 2.12).

Рисунок 2.12 – Функции выделения наборов rgb-каналов

Далее выведем также функцию для наглядного отображения маски *show\_mask* и нормализации яркости и контрастности изображения *show\_img\_norm* (рис. 2.13).

Рисунок 2.13 – Функции show\_mask и show\_img\_norm

Посмотрим на результат работы созданных функций. Будем выводить рядами старый снимок, затем новый снимок и маску с выделенными областями. Несколько примеров полученных изображений представлены на рисунке 2.14.





Рисунок 2.14 – Визуализация снимков

После того, как мы выделили обычную rgb-картинку из спутникового снимка, стоит учесть, что данные в невидимых спектрах при этом тоже очень полезны. Например, в инфракрасном диапазоне легче увидеть пожары (данные о теплоте). Кроме этого, можно использовать разные комбинации каналов для выявления определенных особенностей изображения. Общеизвестна такая комбинация и для выделения данных о растительности, основанная на особенностях отражения света от листьев.

Для задачи выявления нарушений в состоянии лесного покрова будем использовать нормализованный разностный индекс растительности (NDVI). Высокие значения NDVI соответствуют областям, которые больше отражают ближний инфракрасный спектр, что соответствует более густой и здоровой растительности.

где NIR - отражение в ближней инфракрасной области спектра,

RED - отражение в красной области спектра.

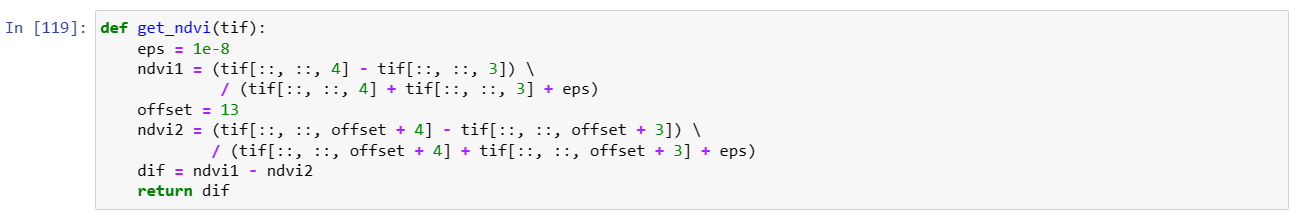
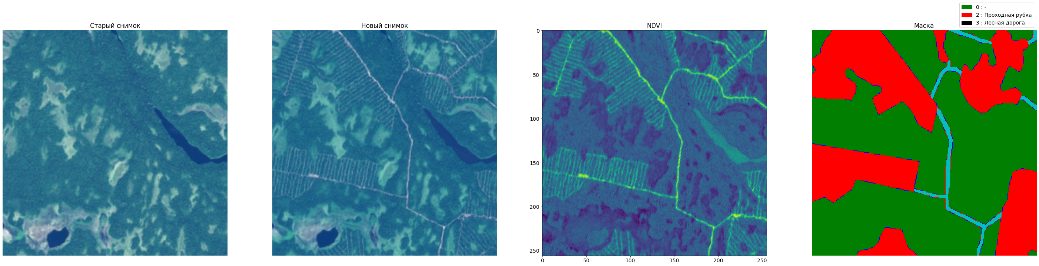
У спутников Sentinel 2, которые сделали снимки, NIR и Red соответствуют каналам B8 и B4 (таблица 1). Выделим функцию get\_ndvi и посмотрим на снимки в этой комбинации каналов (рис. 2.15).

Рисунок 2.15 – Функция добавления каналов NDVI

 Теперь вновь выведем изображения. Будем выводить также сначала пару снимков со спутника, потом вид совмещенного старого и нового снимков с индексом NDVI, после которого будет выводиться маска (рис. 2.16).

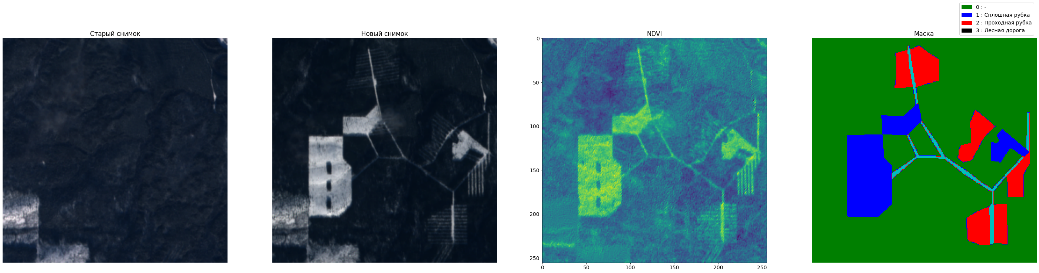


Рисунок 2.16 – Визуализация снимков с NDVI

В дальнейшем при обучении добавим NDVI к первоначальному набору каналов с помощью функции *add\_ndvis* (рис. 2.17). Увеличив количество признаков (26 каналов + 1 ndvi), мы добьемся более точных результатов.

Рисунок 2.17 – Функция add\_ndvis

1. **ОПИСАНИЕ МЕТОДА ОБУЧЕНИЯ**

В этой части работы мы обучим модели бинарной классификации областей леса на основе уже ранее подготовленных обучающей и тестовой выборок и оценим полученные модели. Для решения задачи бинарной классификации и определения наличия и отсутствия повреждения будем использовать алгоритм RandomForest.

* 1. **Применение алгоритма RandomForest**

Random Forest или «случайный лес» – это алгоритм машинного обучения, который используется для решения задач классификации и регрессии. Он является расширением алгоритма решающих деревьев, который использует ансамбль деревьев для улучшения качества классификации или регрессии.

Суть алгоритма заключается в том, что он создает множество решающих деревьев и использует их для предсказания классов объектов. Каждое дерево строится на случайном подмножестве обучающих данных и случайном подмножестве признаков. В результате, каждое дерево в ансамбле получается немного разным, что позволяет уменьшить эффект переобучения и повысить качество предсказаний [7].

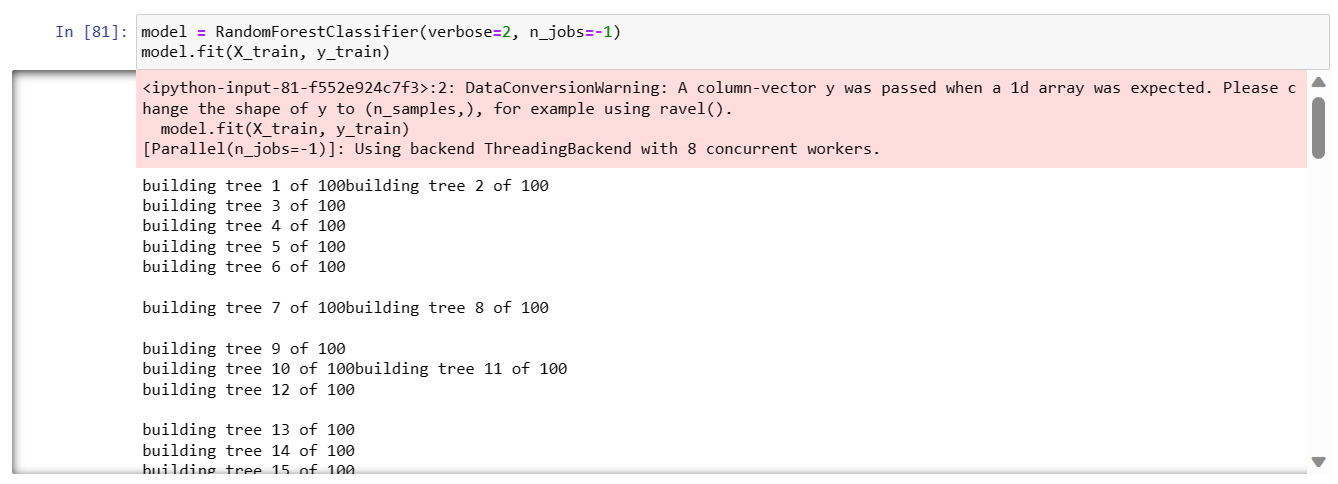
С помощью метода RandomForest была проведена бинарная классификация (рис. 3.1)

Рисунок 3.1 – Алгоритм RandomForest

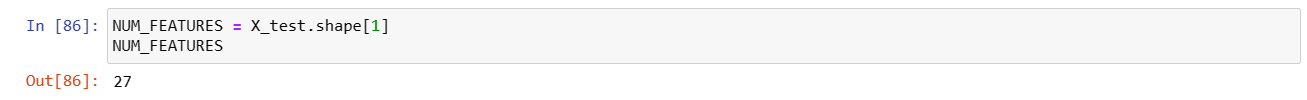
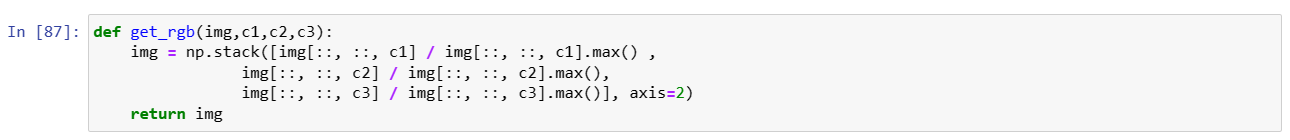
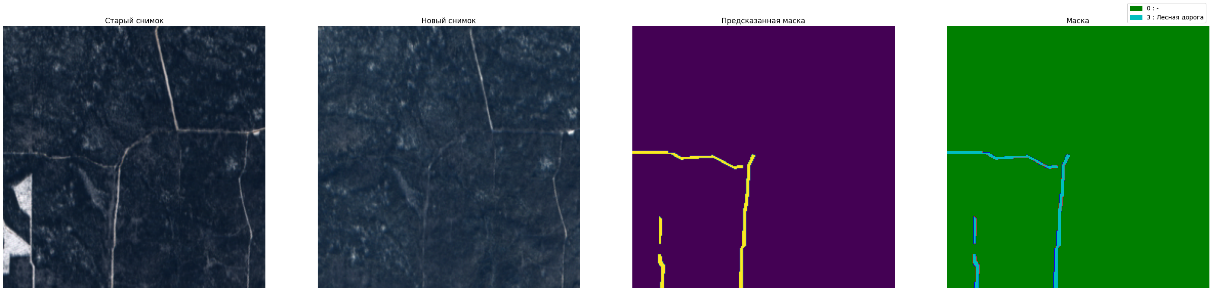
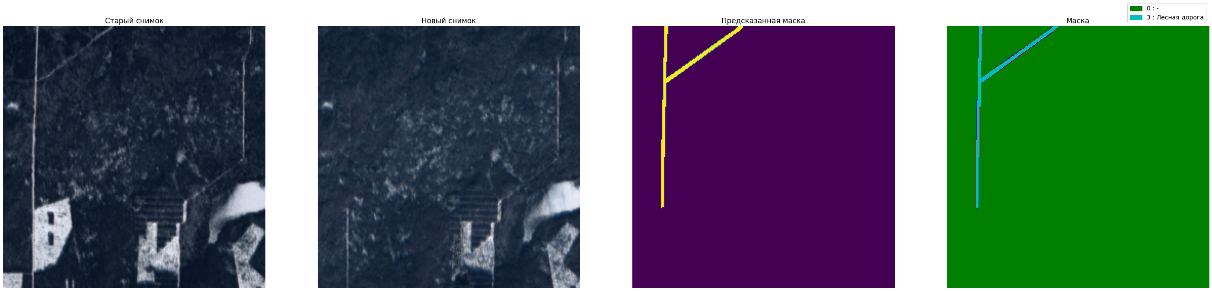
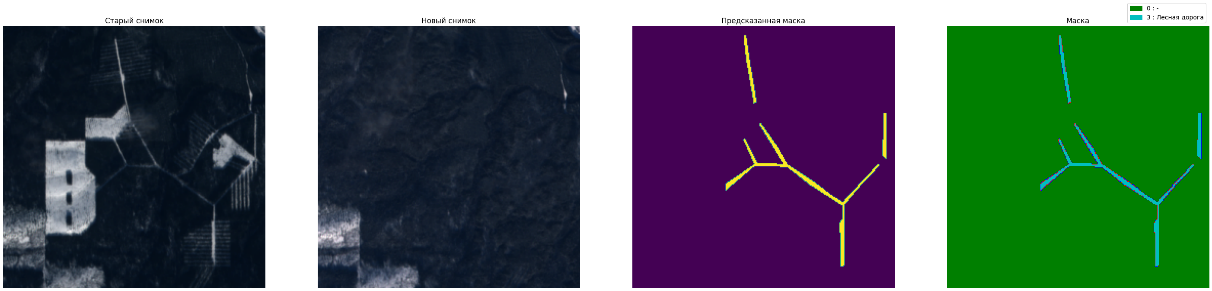
Для того, чтобы посмотреть на предсказанные маски, вновь переформируем сплошной массив 0 и 1 в картинки с помощью уже использованного reshape. Также нам придётся использовать новое get\_rgb (рис.3.2), ведь в отобранных тестовых данных (X\_test) номера каналов смещены на 1, т.к. нет нулевого канала с маской (рис. 3.3).

Рисунок 3.2 – Смещение каналов

Рисунок 3.3 – Новая функция get\_rgb

Посмотрим на результаты работы обученной модели с помощью алгоритма RandomForest. Будем выводить сначала пару снимков старый-новый, затем предсказанную моделью маску, после изначальную маску (рис. 3.4).

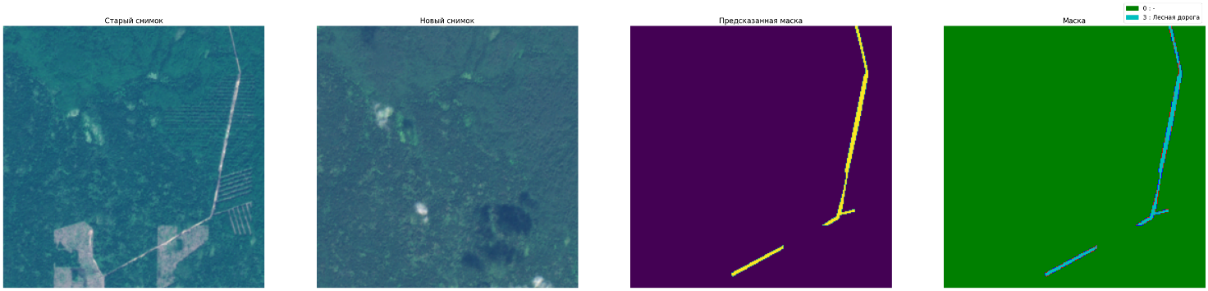


Рисунок 3.4 – Результат обученной модели

Из полученных изображений можно сказать, что созданная модель достаточно точно определяет необходимые области леса. Также для наглядности сравним результат с изначально построенной маской. Из рисунка 3.4 видно, что модель действительно с некоторой точностью определяет искомые нами сегменты, но для достоверности, стоит рассчитать степень точности обученной модели.

* 1. **Оценка обученной модели**

Для оценки эффективности модели рассчитаем Dice коэффициент, который показывает попиксельное соотношение между прогнозируемой маской и соответствующей ей истиной. Коэффициент Dice — это двойная площадь их пересечения, деленная на общее количество пикселей в обоих изображениях.

где и - мощности двух наборов, т. е. количество элементов в каждом наборе.

Dice принимает значения от 0 до 1 (полное совпадение).

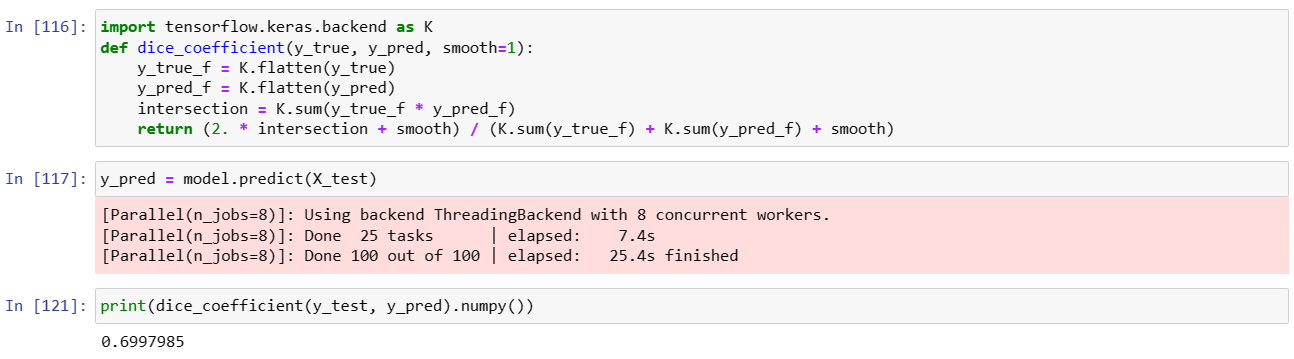
Реализуем функцию для расчета коэффициента Дайса (рис. 3.5).

Рисунок 3.5 – Коэффициент Дайса

В итоге мы получили коэффициент приблизительно равный 0.7, что является неплохим результатом работы модели, однако видно, что найденные как в изначальной, так и в предсказанной моделью масках дороги получаются прерывистыми, не соответствующими исходным снимкам. Так происходит из-за наложения друг на друга других областей леса. Попробуем исправить ситуацию, применив фильтры.

* 1. **Фильтр Габора**

Фильтры используются для выделения разных низкоуровневых признаков из исходного изображения. Есть фильтры, выделяющие границы, есть размывающие или зашумляющие изображение фильтры. У нас сейчас уже довольно много признаков (26 каналов + ndvi), но можем рискнуть добавить ещё. Для работы реализуем фильтр Габора.

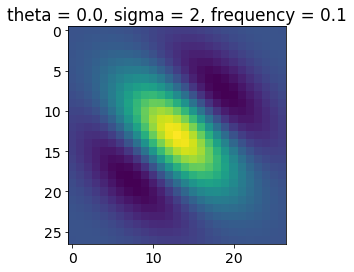
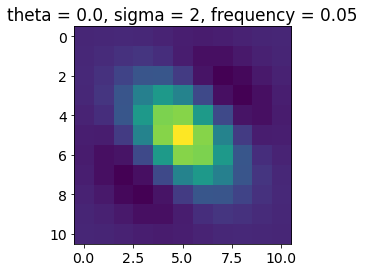
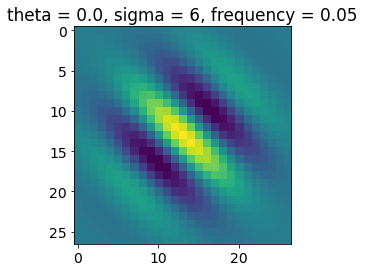
Фильтр Габора — это линейный полосовой фильтр, т.е. он пропускает частоты в определенной полосе и ослабляет другие частоты вне этой полосы. Несколько фильтров можно определить, изменив параметры \lambda, \theta, \sigma, \psi и \gamma. Банк фильтров Габора представляет собой набор фильтров Габора с различными параметрами. Добавим функции для реализации фильтра Габора и посмотрим на то, как его параметры виляют на изображение (рис. 3.6).

Рисунок 3.6 – Фильтр Габора

Введем некоторые вспомогательные функции для применения 16 фильтров к ndvi-каналу и добавления к исходным (рис. 3.7).

Рисунок 3.7 – Функции для фильтра

Посмотрим на изменение количества каналов у получившегося набора данных (рис. 3.8).

Рисунок 3.8 – Количество добавленных каналов

Количество каналов теперь 42, из которых 27 исходных каналов и 16 фильтров на NDVI. Теперь посмотрим на изменение изображений, визуализировав пары старых-новых снимков и изображения с NDVI (рис. 3.9).

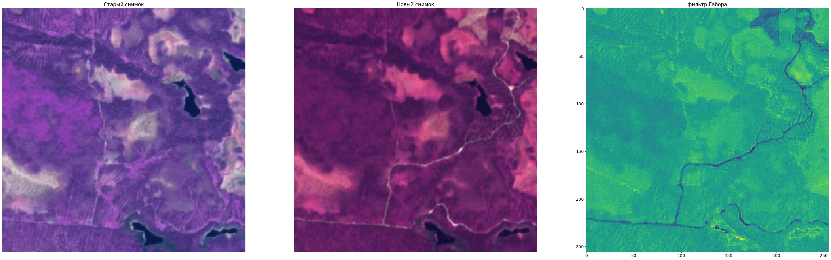
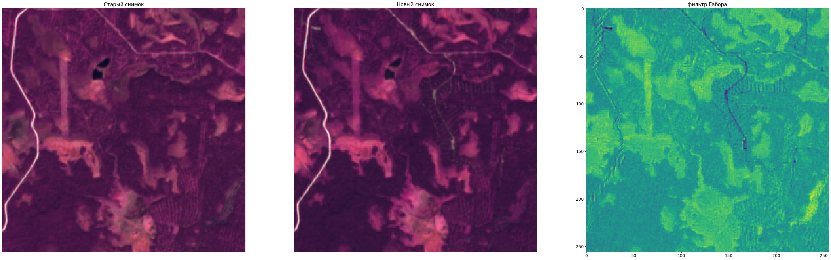
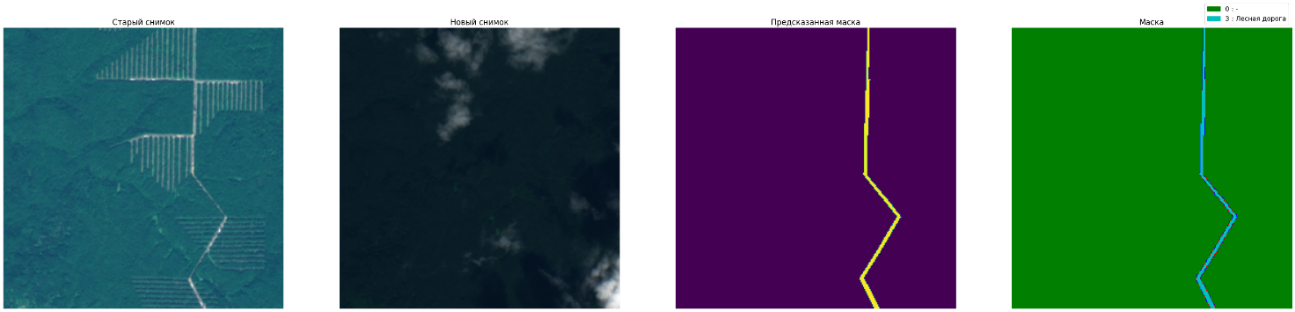
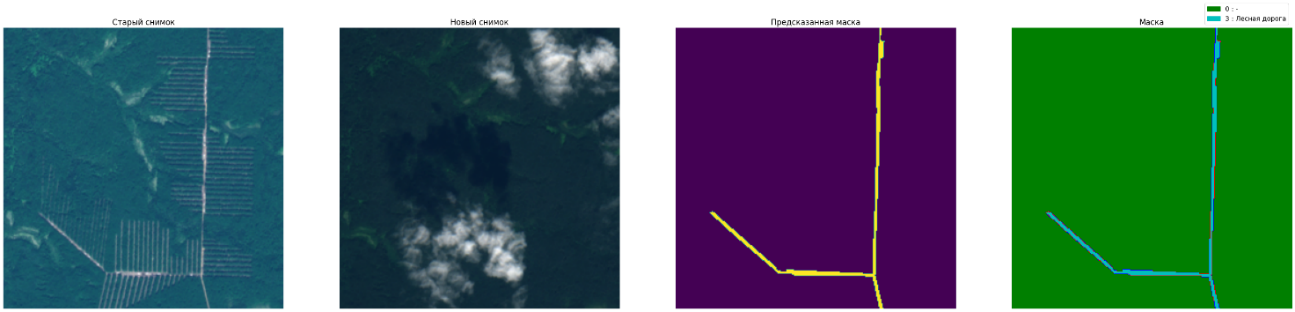


Рисунок 3.9 – Визуализация изображений с 26 каналами

Видно, что увеличение каналов привело к изменению исходных снимков. Изменение таких параметров как длина волны меняет ширину полос функции Габора (уменьшение длины волны дает более тонкие полосы), параметр \sigma регулирует количество полос пропускания, а \gamma управляет высотой функции Габора. Для очень высокого \gamma высота становится очень маленькой, а для очень малого значения - высота становится довольно большой.

На основе новых данных обучим модель и посмотрим на результаты работы программы через созданные ранее функции визуализации (рис. 3.10)





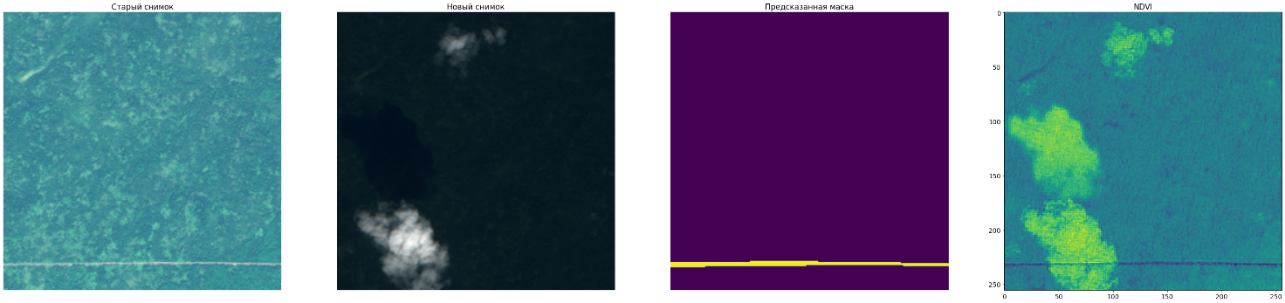


Рисунок 3.10 – Результат обученной модели с фильтром Габора

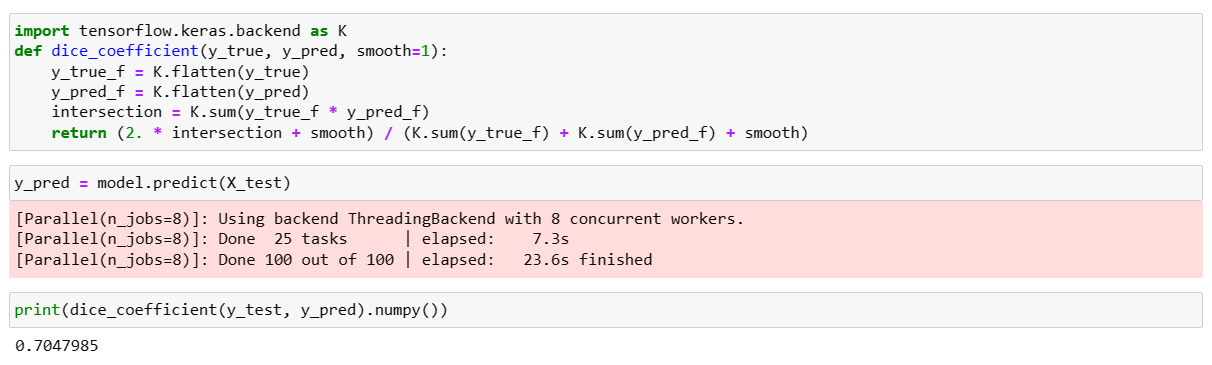
В результате были получены удачные варианты нахождения сегментов, где лесные дороги не обрываются и более точно соответствуют исходным снимкам. Посмотрим, изменился ли коэффициент Дайса, для сравнения с ранее построенной моделью без фильтра Габора (рис. 3.11).

Рисунок 3.11 – Коэффициент Дайса для новой модели

Видно, что коэффициент стал немного больше, что говорит о большой точности полученной модели. Из-за того, что сегмент с лесными дорогами хуже распознается на исходных снимках в отличие от других классов, недостаточное количество информации приводило к не самым удачным результат в первой модели, но при использовании фильтров видимость дорог стала лучше, соответственно и результаты второй модели стали точнее.

# **ЗАКЛЮЧЕНИЕ**

В результате выполнения данной научно-исследовательской работы были выполнены следующие задачи:

1. Были изучены и описаны полученные данные. Определены формат снимков, их число, размерность, расположение в директориях, а также каналы изображений.
2. Полученные данные были визуализированы в парах снимков старый-новый в формате rgb. Был изучен и выведен индекс NDVI, и построена маска изменение области
3. Были сформированы обучающие и тестовые выборки, содержащие повреждения заданного вариантом класса, чтобы выполнить необходимую классификацию.
4. Был применен алгоритм RandomForest и на его основе сформирована модель, обученная бинарной классификации. Проведена оценка полученной модели с помощью коэффициента Дайса. Был улучшен результат работы модели путем применения фильтра Габора.

# **СПИСОК ЛИТЕРАТУРЫ**

Abdi A. M. Land cover and land use classification performance of machine

learning algorithms in a boreal landscape using Sentinel-2 data // GIScience Remote Sens.

Taylor & Francis. 2020. Vol. 57, № 1. P. 1–20

1. Abdi A. M. Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data // GIScience Remote Sens. Taylor & Francis. 2020. Vol. 57, № 1. P. 1–20.
2. Carranza-García M., García-Gutiérrez J., Riquelme JC. A framework for evaluating land use and land cover classification using convolutional neural networks // Remote Sensing, 2019. 11(30). 274.
3. Табунщик В.А., Петлюкова Е.А., Хитрин М.О. Труды Карадагской научной станции им. Т.И. Вяземского - Природного щаповедника РАН // Применение спутниковых снимков Sentinel-2 для анализа земель используемых в сельском хозяйстве (на примере раздольненского района республики Крым), 2018. №1 (5). С. 43–57.
4. Бычков И. В., Ружников Г. М., Федоров Р. К., Попова А. К., Авраменко Ю. В. Классификация космоснимков Sentinel-2 Байкальской природной территории // Компьютерная оптика. 2022. Т. 46, № 1. С. 90–96.
5. Курбанов Э. А., Воробьев О. Н., Меньшиков С. А., Смирнова Л. Н. Распознавание лесных насаждений и доминирующих древесных пород Пензенской области по данным спутника Sentinel-2 // Современные проблемы дистанционного зондирования Земли из космоса. 2018. Т. 15. №5. С. 154–166
6. «NumPy» [Электронный ресурс] – Режим доступа: [https://numpy.org/](https://numpy.org/doc/stable/user/basics.rec.html)
7. «Random Forest, метод главных компонент и оптимизация гиперпараметров: пример решения задачи классификации на Python» [Электронный ресурс] – Режим доступа: <https://habr.com/ru/companies/ruvds/articles/488342/>