

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение  
высшего образования

«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ  
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

**ЗАДАЧА КЛАССИФИКАЦИИ ТЕРРИТОРИИ ВОДОХРАНИЛИЩА  
НА РЕЧНЫЕ И ОЗЕРНЫЕ УЧАСТКИ ПО ДАННЫМ  
ДИСТАНЦИОННОГО ЗОНДИРОВАНИЯ ЗЕМЛИ (ДЗЗ)  
АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ**

студентки 2 курса 273 группы

направления 02.04.03 Математическое обеспечение и администрирование  
информационных систем

факультета компьютерных наук и информационных технологий

Ганюшкиной Анны Вячеславовны

Научный руководитель:

зав. кафедрой ИиП, к.ф.-м.н., доцент \_\_\_\_\_ Огнева М.В.

подпись, дата

Зав. кафедрой:

к.ф.-м.н., доцент \_\_\_\_\_ Огнева М.В.

подпись, дата

Саратов 2026

## ВВЕДЕНИЕ

**Актуальность темы.** Данные дистанционного зондирования Земли (ДЗЗ) выступают одним из фундаментальных источников информации для современных геоэкологических и картографических исследований. Они обеспечивают возможность системного картографирования ландшафтных структур и объектов инфраструктуры, эффективного управления природными ресурсами и всестороннего анализа процессов изменения окружающей среды. С развитием технологий получения изображений и совершенствования методов автоматизированного дешифрования, материалы космической съемки высокого разрешения приобрели особую значимость как оперативный, надежный и эффективный инструмент мониторинга состояния и динамических изменений в землепользовании. Введение ДДЗ в анализ территорий способствует повышению точности картографических продуктов, ускорению обновления данных и расширению возможностей пространственного моделирования изменений. Применение дешифрирования разновременных космических снимков в целях картографирования структуры землепользования является весьма востребованным, так как оно обеспечивает качественное изучение изменений исследуемого объекта для дальнейшего создания баз данных [1].

Наиболее рациональным и эффективным методом хранения и обработки данных природных территориальных систем является метод геоинформационного картографирования. В его основе лежит использование специального программного обеспечения – геоинформационных систем (ГИС), необходимых для сбора, хранения, обработки и визуализации пространственно-координатных данных [2]. ГИС-технологии – это важный атрибут в жизни современного общества. Они находят применение как в глобальных вопросах: загрязнение территорий, сокращение лесных угодий, последствия катастроф природного и антропогенного генезиса, перепроизводство сельскохозяйственной продукции, так и в решении более повседневных задач: поиск дома по его адресу, подбор наиболее

оптимального пути, логистические маршрутные построения, проведение водопровода или линий электропередач.

Несмотря на достаточно большое практическое значение данных, полученных с помощью ГИС-технологий, разметка и анализ каждого изображения требует непосредственного участия специалиста, что сильно сокращает эффективность их работы.

Актуальность темы заключается в возможности обрабатывать большой объем данных из одного или нескольких источников в режиме настоящего времени. Кроме того, создание приложения для обработки и извлечения данных изображения, полученного с помощью кроссплатформенной геоинформационной системы QGIS, позволит значительно сократить трудозатраты географов при классификации и обработке ДЗЗ и поможет обеспечить регулярное обновление данных по большим территориям.

**Цель магистерской работы** – создание приложения для автоматической классификации территории водохранилища на речные и озерные участки по данным дистанционного зондирования земли (ДЗЗ) методами машинного обучения.

Поставленная цель определила **следующие задачи**:

1. Исследование существующих решений задачи классификации территории по данным дистанционного зондирования Земли.
2. Рассмотрение методов машинного обучения для решения задачи классификации табличных данных.
3. Рассмотрение методов машинного обучения, позволяющих повысить точность классификации при дисбалансе классов.
4. Изучение способов оценки качества полученных результатов.
5. Обработка космических снимков и извлечение признаков.
6. Реализация выбранных алгоритмов.
7. Анализ результатов и визуализация ошибок моделей, построенных на предыдущем этапе.
8. Доработка классификаторов, реализация альтернативных

алгоритмов, анализ точности.

9. Разработка и интеграция в геоинформационную систему QGIS модуля для обработки растров и классификации с помощью обученных ранее моделей.

10. Апробация разработанного программного модуля.

**Методологические основы** использования ГИС-технологий при классификации и исследования территорий методами машинного обучения представлены в работах Непоклонова, Наумова, Кутявиной, Рутман, Мирненко, Вирьясепутра.

**В теоретической части** приведен обзор существующих решений, изложена общая информация о методах машинного обучения для решения задачи классификации и приведен используемый инструментарий.

**В практической части** работы рассмотрены реализация и сравнительный анализ применения методов машинного обучения для классификации табличных данных, а также реализация программного модуля QGIS.

**Структура и объём работы.** Магистерская работа состоит из введения, четырех разделов, заключения, списка использованных источников и одиннадцати приложений. Общий объем работы – 137 страниц, из них 88 страниц – основное содержание, включая 29 рисунков и 4 таблицы, список использованных источников информации – 62 наименования.

## **КРАТКОЕСОДЕРЖАНИЕ РАБОТЫ**

**Первый раздел «Обзор существующих решений»** посвящен анализу существующих подходов и программно-технических средств, применяемых при дистанционном мониторинге водных объектов и классификации территорий по космическим снимкам.

В подразделе «Исследование методов машинного обучения» проанализированы исследования, посвященные изучению процессов эвтрофирования водоемов и применяемым методам их оценки, в частности

методы с расчетом спектральных индексов и построением индексных карт. Рассмотрены работы [3], [4], [5] и другие по прогнозированию качества воды с применением методов машинного обучения (решающие деревья, случайный лес, наивный Байес, SVM, KNN, XGBoost, градиентный бустинг, нейронные сети), включающие сравнительный анализ и оценку точности моделей на реальных данных. Описаны исследования, использующие спутниковые данные и спектральные индексы.

В результате, существуют решения, направленные на оценку качества воды и степень эвтрофирования водоемов, а также классификацию водных объектов из спутниковых снимков. Однако еще не было представлено комплексного решения, позволившего бы по космоснимку разметить территории на речные и озерные участки, а также наглядно показать значения ключевых признаков, влияющих на степень эвтрофирования водоемов, например, температуры.

В подразделе «Обзор геоинформационных систем» изложена информация об основных геоинформационных системах (ArcGIS, GEE, QGIS, GRASS, ILWIS), приведены их особенности и недостатки.

Географические информационные системы (ГИС) представляют собой ключевой программный инструментарий геопространственных технологий, используемый для сбора, хранения, обработки, анализа и представления различных типов пространственных и географических данных.

В подразделе «Данные дистанционного зондирования Земли» приведен обзор данных дистанционного зондирования (ДЗЗ) как ключевого источника геопространственной информации: понятие ДЗЗ, принципы получения данных, преимущества спутниковых наблюдений и роль ДЗЗ в классификации и картографировании. Рассмотрены основные форматы данных (GeoTIFF, NetCDF, Shapefile) для хранения, предобработки и интеграции растровых и векторных данных.

Дистанционное зондирование Земли (космический мониторинг) – это систематическое получение информации о состоянии земной поверхности с

космических аппаратов.

Пространственные данные (геопространственные данные (сведения), геоданные) представляют собой сведения о форме пространственных объектов, их местоположении и свойствах, данные о процессах и явлениях, в том числе представленные с использованием координат.

В геоинформационных системах обычно выделяют два основных типа геопространственных данных: векторные и растровые. Векторные данные описывают объекты в виде точек, линий и полигонов. Растровые данные представляют территорию в виде матрицы пикселей, каждый из которых содержит значение яркости или отражательной способности в определенном спектральном канале.

**Второй раздел «Методы машинного обучения для решения задачи классификации»** посвящен обзору алгоритмов и подходов, применяемых для классификации табличных данных, методам обучения для сложных случаев и способам оценки качества классификаторов.

В подразделе «Классические методы решения задачи классификации» изложены классические методы классификации: метод k-ближайших соседей (KNN), дерево решений и ансамблевые методы (стекинг, случайный лес, градиентный бустинг). Описаны принципы работы каждого метода, достоинства и ограничения.

В подразделе «Методы машинного обучения для сложных случаев классификации табличных данных» рассмотрены подходы для трудных случаев классификации: кластеризация (метод k-средних, BIRCH, агломеративная кластеризация, спектральная кластеризация) и нейронные сети (архитектура, функции активации и многослойный перцептрон).

В подразделе «Способы оценки качества классификации» представлены способы оценки качества классификаторов: матрица ошибок, численные метрики (accuracy, sensitivity, specificity, precision, recall, F1-мера) и графическая (ROC-AUC).

Для корректной оценки моделей необходимо использовать несколько

метрик, ориентируясь на решаемую задачу.

**Третий раздел «Используемый инструментарий»** посвящен выбору и обоснованию средств программной реализации исследований по обработке растровых и векторных пространственных данных, обучению и сравнению моделей машинного обучения, а также разработке модуля классификации в геоинформационной системе.

В разделе приведено описание аппаратной и программной среды разработки. Все рассмотренные на практике алгоритмы и программный модуль были реализованы на языке Python. Необходимые для классификации подсчеты и эксперименты велись в настольной версии системы QGIS.

**Четвертый раздел «Практическая часть»** посвящен реализации рассмотренных алгоритмов машинного обучения, их сравнительному анализу и разработке программного модуля для геоинформационной системы QGIS.

В подразделе «Предварительная подготовка данных в системе QGIS» подробно описывается процесс вычисления признаков на основе полученных растров отражения в синем, зеленом, красном и ближайшем инфракрасном каналах спектра и метаданных.

В подразделе «Сравнительный анализ методов классификации» приводятся результаты тестирования алгоритмов для разметки территорий водохранилища по типам.

В качестве данных был использован датасет, содержащий 351628 точек, принадлежащих одному из четырех классов: коренной берег, суша в ложе, открытая вода (озерный тип) и вода в протоках (речной тип).

В результате исследования распределения классов была обнаружена их существенная несбалансированность. Соотношение количество точек, принадлежащих классам, равно: 0 – 1,94%, 1 – 32,8%, 2 – 53,88%, 3 – 11,38%.

Имеющиеся данные были разбиты на несколько выборок, соответствующие решаемым задачам обучения моделей, а именно:

- 1) Выборка, которая содержит все 4 типа.
- 2) Выборка, которая содержит типы 1, 2 и 3.

- 3) Выборка, которая содержит тип 0 и тройку типов 1, 2, 3.
- 4) Выборка, которая содержит тип 1 и пару типов 2, 3.
- 5) Выборка, которая содержит типы 2 и 3.
- 6) Выборка, которая содержит пары типов 0, 1 и 2, 3.

Сначала была реализована классификация всех четырех типов, определенных в датасете. В результате очень небольшое количество объектов, принадлежащих типу 0, по сравнению с другими типами сказывается на качестве моделей. Большая часть объектов этого класса определяется к типу 1. Лучшие результаты показали модели, использующие CatBoost, LGBMClassifier и балансировку методом взвешивания классов для случайного леса.

Основной же интерес для исследования в данной работе представляло отделение открытой воды (озерного типа) от воды в протоках, островах и т.п. (речной тип). Для решения этой задачи были построены модели машинного обучения, основанные на ранее рассмотренных алгоритмах. Обучение происходило на переопределенном датасете. В него были добавлены только объекты, принадлежащие типам 2 и 3, причем им были назначены новые типы: для класса 2 – 0, для класса 3 – 1. Лучшие результаты показали CatBoost и случайный лес с accuracy в 90,7%. Однако CatBoost выиграл в точности: 0.94 для класса 2 и 0.76 для класса 3.

Помимо основной задачи отделения озерных и вечных участков воды, бинарные классификации могут решить ряд других задач, таких как: отделить коренной берег от остальных типов (класс 0 против классов 1, 2 и 3), отделить острова и полуострова от воды (класс 1 против классов 2 и 3), отделить сушу от воды (классы 0 и 1 против классов 2 и 3). Объединение моделей таких классификаторов может решить основную задачу для любых входных данных. Рассмотренные бинарные классификаторы позволили определить 2 последовательных конвейера.

Принцип работы первого:

- 1) Отделить класс 0 от всех остальных.

- 2) Отделить класс 1 от класса 2 и 3.
- 3) Отделить классы 2 и 3 между собой.

Принцип работы второго:

- 1) Отделить классы 0 и 1 от классов 2 и 3.
- 2) Отделить классы 2 и 3 между собой

На вход конвейерам подается массив точек всех четырех типов. На выходе получается набор только тех точек, которым был предсказан один из классов 2 и 3.

В итоге, в финальный набор попали около 2 тысяч объектов, принадлежащих изначальным классам 0 и 1 – это около 3% от числа всего набора. Конвейеры показали 89% правильности результата. Точность предсказания класса 2 составила 0.94, класса 3 – 0.63 и 0.65 для первого и второго конвейеров соответственно, что выше этого же показателя для многоклассовых классификаторов, но ниже, чем у бинарного.

Реализованные алгоритмы кластеризации, за исключением метода k-средних, не подходят для решения поставленной задачи. Метод k-средних может быть использован вместе с другими моделями классификаторов. Но при использовании кластеризации для решения задачи классификации есть сложность в определении к какому классу какой кластер был предположительно соотнесен.

Для моделей нейронных сетей с одной и 4 скрытыми слоями в качестве функции активации использовался ReLU, для выходного слоя – сигмоида. Модель с 4 скрытыми слоями показала большую точность и полноту для класса 1 (речной тип), по сравнению с ранее реализованными классическими моделями, что важно ввиду несбалансированности классов в выборке.

Из результатов визуализации ошибок классификаторов можно сделать вывод, что наибольшую трудность для классификаторов представляют участки речного типа, ограниченные островами, но не имеющие выраженных признаков этого типа, а также некоторые озерные участки вблизи островов. Данные участки расположены на границе типов и могут вызвать сложности

при разметке даже у специалиста.

В подразделе «Разработка программного модуля в QGIS» описываются функциональные возможности плагина, его архитектура, процесс использования и апробация на реальных данных.

В качестве пользовательского интерфейса для обработки растров и классификации с помощью обученных ранее моделей был разработан плагин для геоинформационной системы QGIS. Его реализация позволяет проводить все этапы подготовки данных для классификации и дальнейшее предсказание классов автоматически, используя возможности среды ГИС.

В процессе тестирования функциональности плагина для обработки и классификации геопространственных данных территории Волгоградского водохранилища в центральном районе Саратовской области было установлено, что полученные результаты соответствуют первоначальным требованиям и ожиданиям. Основная задача алгоритмов заключалась в корректной обработке данных из выбранного источника и их последующей визуализации в виде цветных точек по заданным. Данная задача была успешно выполнена.

## **ЗАКЛЮЧЕНИЕ**

В ходе работы были рассмотрены существующие решения. Несмотря на наличие готовых решений классификации водоемов, интерес представляет их классификация именно с помощью фиксации отражения водной поверхности в разных каналах спектра автоматизировано выделять участки условно «речного» типа и условно «озерного» типа, используя при этом данные, обработанные и организованные в уже известные спектральные индексы, а также растр поверхностных температур.

На теоретическом уровне были рассмотрены такие методы машинного обучения, решающие задачу классификации, как метод k-ближайших соседей, деревья решений, случайный лес, градиентный бустинг и стекинг, метод k средних, BIRCH, агломеративная кластеризация, многослойный

перцептрон. Также были рассмотрены способы оценки качества классификации, среди которых: accuracy, sensitivity и specificity, precision и recall, f1-мера, ROC-AUC.

На практике была проведена обработка космического снимка в системе работы с геопространственной информацией QGIS. Для датасета, содержащего данные о точках, принадлежащий четырем классам: коренной берег, суша в ложе, речной и озерный тип – были реализованы различные модели многоклассовой и бинарной классификации, использующие рассмотренные в теоретической части методы, и проведен их анализ. Небольшое количество объектов, принадлежащих классам 0 и 3, привело к их плохой распознаваемости. Кроме того, были реализованы два конвейера бинарных классификаторов. В результате лучшие результаты показали модель, использующая классификатор из библиотеки CatBoost, и модель многослойного перцептрона, построенная с помощью модуля keras библиотеки TensorFlow. Визуализация и анализ ошибок моделей классификации показали, что большинство из них связаны с определением достаточно обширных речных участков, окруженных островами.

Для практического применения моделей был разработан плагин для геоинформационной системы QGIS, реализующий все необходимые этапы обработки спутниковых снимков, начиная от вычисления спектральных индексов из растровых слоев до получения векторного слоя с размеченными классами.

Результаты исследования были представлены на Студенческой научной конференции факультета компьютерных наук и информационных технологий СГУ.

Таким образом, все поставленные задачи были выполнены и цель достигнута.

**Отдельные части магистерской работы были представлены на конференции:**

**Основные источники информации:**

1. Непоклонов, В. Б., Хабаров, Д. А., Хабарова, И. А. Применение геоинформационных технологий при исследовании изменений структуры землепользования территорий по материалам многозональной космической съемки // Вестник СГУГиТ (Сибирского государственного университета геосистем и технологий) – Т. 25, № 2 – 2020 – С. 151-169.
2. Наумова, Т. А. Использование ГИС-технологий в решении экологических проблем // Научно-образовательный журнал для студентов и преподавателей "StudNet" – Т. 3, № 3 – 2020 – С. 312-318.
3. Кутявина, Т. И., Рутман, В. В., Ашихмина, Т. Я. Применение пространственного геоинформационного анализа по материалам космоснимков для проведения мониторинга состояния эвтрофированных водоёмов Кировской области // Академику Л.С. Бергу – 145 лет: Сборник научных статей. Бендеры. 2021. Т. I. С. 150-152.
4. Мирненко, Э.И. Особенности эвтрофирования Нижнекальмиусского водохранилища // Проблемы экологии и охраны природы техногенного региона, № 3-4 – 2019 – С. 24-30.
5. Wiryaseputra, M. Water Quality Prediction Using Machine Learning Classification Algorithm // International Journal of Scientific & Engineering Research – Vol. 8. No. 9. – October 2022 – P. 4.